Vision artificielle pour les non-voyants : une approche bio-inspirée pour la reconnaissance de formes

Adrien Brilhault

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par

Université Toulouse III Paul Sabatier (UT3 Paul Sabatier)

Discipline ou spécialité

Intelligence Artificielle

Présentée et soutenue par :

Adrien Brilhault

le : 18 Juillet 2014

Titre :

Vision artificielle pour les non-voyants : une approche bio-inspirée pour la reconnaissance de formes

---

Ecole doctorale

Mathématiques Informatique Télécommunications (MITT)

Unité(s) de recherche :

Centre de Recherche Cerveau & Cognition – UMR 5549
Institut de Recherche en Informatique de Toulouse – UMR 5505

Directeur(s) de Thèse :

Christophe Jouffrais
Simon Thorpe

Rapporteurs :

Michel Paindavoine
Evelyne Klinger

Autre(s) membre(s) du jury :

Jean-Pierre Jessel
Guido Bologna
Laurent Perrinet

Résumé

Vision artificielle pour les non-voyants :
une approche bio-inspirée pour la reconnaissance de formes

La déficience visuelle touche aujourd'hui plus de 315 millions de personnes à travers le monde, un chiffre qui pourrait doubler d'ici 2030 du fait du vieillissement de la population. De par la diversité de ses causes, le nombre de personnes atteintes, et ses conséquences sur la qualité de vie, cette affection fait partie des problèmes de santé d'importance majeure. Les deux grandes approches holistiques pour compenser la perte ou l'absence de vision sont les systèmes de substitution sensorielle, restituant l'information visuelle par l'intermédiaire d'une autre modalité sensorielle (généralement l'audition ou le toucher), et les neuroprothèses visuelles. Ces dernières reproduisent à la surface du relais visuel implanté les images acquises par une caméra, en respectant leur configuration spatiale, un pixel correspondant à une électrode. Malheureusement, les neuroprothèses actuelles souffrent encore d'une perte de résolution trop importante, puisqu'une image ne sera restituée que par une matrice de quelques dizaines de points, rendant ces systèmes inadaptés à une utilisation au quotidien. Ces limitations sont de même nature dans le cas des dispositifs de substitution sensorielle : la quantité d'informations visuelles nécessaire à l'interprétation d'une scène naturelle est bien trop importante par rapport à la résolution de l'interface de restitution (auditive, tactile, ou par micro-stimulation). Ces systèmes se montrent par conséquent inefficaces dans des environnements visuels complexes, et ils ne sont donc qu'extrêmement peu utilisés en dehors des laboratoires de recherches.

Ce constat nous a conduits à proposer dans cette thèse une approche alternative, consistant en un système de suppléance intégrant des méthodes de vision artificielle, afin de prétraiter la scène visuelle, et de ne restituer au non-voyant que les informations extraites pertinentes. Grâce à la reconnaissance de formes en temps réel et à la synthèse de sons spatialisés, ce système permet de restaurer des boucles visuomotrices qui rendent à nouveau possibles certaines fonctions visuelles comme la localisation et la préhension d'objets. La navigation étant une autre tâche critique pour les non-voyants, nous avons également incorporé au dispositif des fonctions de guidage basées sur le positionnement par satellites et sur un système d'information géographique adapté. La trop faible précision de localisation du GPS nous a amenés à développer une nouvelle méthode de

positionnement hybride, combinant les données satellites et inertielles à la reconnaissance de cibles visuelles géolocalisées. L'utilisation de la vision artificielle a ainsi permis d'améliorer les performances de localisation et d'obtenir une erreur moyenne généralement inférieure à 5 mètres, rendant possible le guidage et la navigation d'un piéton non-voyant.

Afin d'améliorer les performances du module de vision artificielle, constituant le cœur du système, nous avons développé et évalué un nouvel algorithme de reconnaissance de formes bio-inspiré multi-résolutions, reposant sur la librairie Spikenet. Celle-ci utilise un codage de l'information visuelle par latence, et des représentations sous forme d'arêtes orientées, telles que celles observées dans le cortex visuel primaire. Par rapport à l'algorithme originel mono-échelle, cette architecture permet de capturer un spectre de fréquences spatiales plus large. Les traitements à faible résolution permettent ainsi d'améliorer la tolérance aux déformations de l'image, alors que les hautes fréquences spatiales, plus discriminantes, maintiennent une précision suffisamment élevée. De par son fonctionnement en plusieurs passes successives, cette nouvelle architecture permet de plus de diminuer les temps de traitement grâce à une première couche rapide, filtrant les objets à rechercher dans la phase suivante à haute résolution, plus coûteuse en temps de calcul.

Mots-clefs : vision artificielle, reconnaissance de formes, systèmes bio-inspirés, déficience visuelle, systèmes de suppléance, localisation.

Abstract

Artificial vision for the Blind:
a bio-inspired approach for objet recognition

More than 315 million people worldwide suffer from visual impairments, with several studies suggesting that this number will double by 2030 due to the ageing of the population. Given the variety of its causes, the volume of people affected, and its consequences on quality of life, visual impairment constitutes a major current health issue. To compensate for the loss of sight, the two main holistic approaches consist of sensorial substitution and neuroprosthetics. Sensorial substitution devices provide visual information through different sensory modalities (i.e. audition or touch). Neuroprostheses reproduce images acquired by a video camera at the surface of the visual structure implanted (retina, LGN, or visual cortex), respecting their spatial configuration: each electrode corresponds to a given pixel. Unfortunately, current implants still suffer from very low resolution: each image is transmitted via a matrix of only a few dozen electrodes, rendering these systems unsuitable for everyday use. Sensory substitution devices are subject to the same limitations: the amount of information needed to process a natural visual scene is far too important in relation to the output interface resolutions (both auditive and tactile, or through microstimulation). Thus the current holistic systems at present are unable to provide sufficient aid in navigating complex visual environments, and are rarely implemented outside the context of laboratory research.

To overcome these obstacles, we propose the use of artificial vision in order to pre-process visual scenes and provide the user with relevant information. We have validated this approach through the development of a novel assistive device for the blind called 'Navig'. Through shape recognition and spatialized sounds synthesis, this system is able to restore visuomotor loops, allowing users to locate and grab objects of interest. With navigation being one of the most challenging tasks for the visually impaired, we also developed guidance features relying on satellite positioning as well as an adapted geographic information system. Given that GPS accuracy in urban areas remains too low to safely guide blind pedestrians, we developed a new positioning method combining GNSS, inertial sensors and the visual detection of geolocalized landmarks. The use of artificial vision succeeded in

reducing the average positioning error, and as a result provides accurate navigational markers to guide visually impaired users.

To enhance the performance of the visual module, a key component of the system, we further developed a novel bio-inspired multi-resolution algorithm for pattern recognition based on the Spikenet library. It uses latency-based coding of visual information, oriented edge representations and several other mechanisms which essentially mimick the activations of the primary visual cortex. Compared to the original monoscale algorithm, our new architecture captures a far broader spectrum of spatial frequencies. Low-resolution processing allows for improved tolerance to image degradations and deformations, while higher and more discriminative frequencies maintain optimal selectivity. Through our cascaded scheme, combining detections at different resolutions, we significantly reduced processing time. Indeed, a first pass is used to filter objects of interest, and only a few candidates are then tested at a higher resolution.

Keywords: Artificial Vision, Object Recognition, Bio-inspired Systems, Visual Impairment, Assistive Devices, Positioning.

Remerciements

L'étude du système visuel humain et sa modélisation constituent une partie importante des travaux que j'ai pu aborder au cours de cette thèse. Ces domaines, entièrement neufs pour moi lorsque j'ai commencé mon doctorat, s'inscrivent pourtant dans la continuité de ma formation universitaire en intelligence artificielle et de mes projets de recherche précédents en vision par ordinateur, qui m'ont permis de faire le pont vers ce domaine riche que sont les neurosciences. Si mon sujet d'étude se concentrait évidement sur les neurosciences visuelles, ces 4 années au laboratoire du CerCo m'ont permis de découvrir un spectre bien plus large des sciences cognitives, notamment des thématiques comme la perception musicale, la synesthésie, la mémoire, le sommeil, la méditation, les troubles psychiatriques, les interfaces cerveau-machines et bien d'autres sujets fascinants. Ces nombreuses découvertes, dont l'intérêt dépasse grandement le simple cadre de mes recherches, ont été très enrichissantes d'un point de vue personnel, en m'apportant une foule de connaissances sur le fonctionnement du cerveau, et de l'Homme d'une façon plus générale. Elles ont été possibles grâce aux nombreuses conférences de qualité organisées par le CerCo, ainsi que d'autres évènements tels que la semaine du cerveau, le Forum des Sciences Cognitives, ou encore les débats et séminaires de l'association Incognu que j'ai intégrée en commençant ma thèse. Je tiens donc à remercier l'ensemble des chercheurs, et évidement Michelle Fabre-Thorpe, l'ancienne directrice du CerCo, ainsi que Simon Thorpe, qui l'a remplacée depuis quelques mois, pour ce cadre de travail très stimulant et pour l'ouverture scientifique qui règne au sein du laboratoire.

En plus d'être l'actuel directeur du laboratoire Cerveau et Cognition, Simon a aussi été mon directeur de thèse, avec Christophe Jouffrais à l'Institut de Recherche en Informatique de Toulouse. Tous deux ont été des encadrants remarquables, aussi bien professionnellement qu'humainement. Je vous remercie infiniment de m'avoir accompagné et guidé dans cette (longue !) aventure. J'ai traversé quelques épreuves difficiles sur le plan personnel durant cette période, et votre soutien a été précieux. Je tiens donc une nouvelle fois à vous dire que votre compréhension et votre aide durant ces moments délicats m'ont touché. Merci encore. Pour revenir aux aspects professionnels, Simon comme Christophe ont une vision très pluridisciplinaire de la recherche, qui est évidente jusque dans leur parcours. Christophe a ainsi commencé par une thèse de neurosciences (impliquant notamment l'enregistrement intracrânien de neurones chez le singe), et travaille désormais autour des axes de la santé et de l'autonomie dans un laboratoire d'informatique, en poursuivant une

démarche intégrant des domaines aussi variés que la psychologie cognitive, les IHM, la conception participative ou encore le développement et l'évaluation de neuroprothèses visuelles. Simon de son côté, a également toujours suivi une approche transversales des sciences du cerveau, collaborant aussi bien avec des biologistes que des mathématiciens, physiciens ou électroniciens. Ce n'est donc pas un hasard qu'ils aient été tous deux membres de la commission interdisciplinaire du CNRS « Cognition, langage, traitement de l'information : systèmes naturels et artificiels » ! Leur ouverture d'esprit, leur enthousiasme et leur créativité ont rendu cette thèse très stimulante.

Je souhaite bien sûr remercier également tous les membres du jury pour avoir accepté de participer à l'évaluation de ce travail de thèse ainsi que les personnes avec qui j'ai collaboré au cours de ce doctorat : Rudy Guyonneau, pour son soutien scientifique au sein de la société Spikenet, Hung Do-Duy pour m'y avoir accueilli, ainsi que Nicolas Guilbaud, Dominique Couthier, et Jong Allegraud pour leur aide ; l'ensemble des partenaires et collègues du projet Navig ; Sébastien Crouzet et Florian Dramas dont les conseils au début de ma thèse ont été importants ; et les stagiaires qui m'ont aidé à mettre en place différentes expériences.

Ces 4 années de doctorat n'auraient pas été les mêmes sans l'atmosphère agréable qui régnait dans les différentes structures où j'ai eu la chance de travailler. Merci donc à tous mes collègues de l'Irit, du CerCo et de Spikenet. J'ai eu la chance d'y rencontrer plein de personnes exceptionnelles. Je pense par exemple à Gabriel, Rodika, Marlène, Federica, Rama, Laetitia, Laia, Romain, Roger, Mehdi, Edward, Thomas, Damien, Tracy, Jake, Claire et tous les autres avec qui j'ai pu passer de très bonnes soirées à discuter, faire la fête, jouer de la musique dans la cave de la coloc Sansou, finir au petit matin dans celle de la dernière chance, enchainer les concerts de jazz, de foro, d'afrobeat ou encore aller faire du wakeboard, du snowboard, et des parties de squash hebdomadaires pour se maintenir en forme.

Parce qu'il y'a aussi la vie en dehors des labos je tiens évidement à saluer ma famille, mon père, Yannick, pour son soutien, Nébia pour sa présence réconfortante, ainsi que Sarah et Othmane. Des pensées également pour ma mère, qui nous a quittée il y'a bientôt dix ans. Merci enfin à mes amis ayant rendu ces années toulousaines si agréables, Camellia, Clément, Pablo, Marc, Florian, Rémi, Régis, Fredo, John, Fabien, Bastien, Stéphanie, Cécile, Marjolaine, Sandra, et ceux que j'oublie. Enfin à Agnès, qui a éclairé cette fin de thèse, et à Mathilde, l'amie de toujours.

Table des matières

Liste des sigles et abréviations

- AUC Area Under Curve
- AVC Acuité visuelle corrigée
- BOVW Bag-Of-Visual-Word
- BOW Bag-Of-Word
- CBIR Content-Based Image Retrieval
- CCD Charge-Coupled Device
- CGL Corps Géniculé Latéral
- CIF Classification Internationale du Fonctionnement
- CIH Classification Internationale du Handicap
- CIM Classification Internationale des Maladies
- CMOS Complementary Metal Oxide Semiconductor
- CS Colliculus Supérieur
- CSD Color Structure Descriptor
- CVS Comma Separated Values
- DET Detection Error Tradeoff
- DMLA Dégénérescences maculaires liées à l'âge
- DSP Digital Signal Processor
- DV Déficient visuel
- EOA Electronic Orientation Aids
- ETA Electronic Travel Aids
- FN (FN) Faux Négatif (False Negative)
- FP (FP) Faux Positif (False Positive)
- GLOH Gradient Location-Orientation Histogram
- GPS Global Positioning System
- HHM Hidden Markov Model
- HID Handicaps, Incapacités, Dépendance
- HOG Histogram of Gradient
- HRIR Head-Related Impulse Response
- HRTF Head-Related Transfer Function
- HSI Hue Saturation Intensity
- HSV Hue Saturation Value
- IAPB International Agency for the Prevention of Blindness
- IHM Interface Homme-Machine
- LBP Local Binary Patterns
- MCC Matthews Correlation Coeficient

- MDC Markov Decision Problem
- OCR Optical Character Recognition
- OMS Organisation Mondiale de la Santé
- PA Potentiel d'action
- PCA Principal Components Analysis
- PDF Point difficile
- PF Point favori
- PI Point de l'itinéraire
- POI Points d'intérêts
- PR Points de repère
- PR Precision Recall
- PSVA Prosthesis for Substitution of Vision by Audition
- PV Point visuel
- RANSAC Random Sample Consensus
- RBF Radial basis function
- RFID Radio Frequency Identification
- RGB Red Green Blue
- RMS Root Mean Square
- ROC Receiver Operating Characteristic
- RP Rétinite Pigmentaire
- SCD Scalable Color Descriptor
- SIFT Scale-Invariant Transform Feature
- SIG Système d'information géographique
- SLAM Simultaneous Localisation and Mapping
- SNV SpikeNet Vision
- SURF Speeded Up Robust Features
- SVM Support Vector Machine
- TDU Tongue Display Unit
- TFD Taux de fausse découverte
- TFP (TNR) Taux de Faux Positifs (True Negative Rate)
- TMS Transcranial Magnetic Stimulation
- TOF Time-of-Flight
- TVS Tactile Vision System
- TVSS Tactile Vision Substitution System
- UPC Universal Product Code
- VN (TN) Vrai Négatif (True Negative)
- VP (TP) Vrai Positif (True Positive)
- VPN Valeur predictive negative

Introduction

La déficience visuelle, de par le nombre de personnes touchées à l'échelle mondiale et ses conséquences sur la qualité de vie, fait partie des problèmes de santé d'importance majeure. L'Organisation Mondiale de la Santé (OMS) recensait en effet près de 314 millions de déficients visuels (DV) à travers le monde en 2002, parmi lesquels plus de 45 millions d'aveugles [World Health Organization, 2005]. En France, d'après l'enquête HID[1] réalisée en 2005 [Sander et al., 2005], leur nombre s'élèverait à 1,7 millions (dont 207 000 aveugles).

Les causes de malvoyance et de cécité sont nombreuses. Si les glaucomes, cataractes, dégénérescences maculaires, rétinopathies diabétiques, trachomes, onchocercoses ou encore les xérophtalmies comptent parmi les plus fréquentes, l'épidémiologie varie grandement d'une région à une autre [Thylefors et al., 1995]. La difficulté d'accès aux soins dans les pays en voie de développement explique par exemple le nombre d'affections visuelles particulièrement élevé sur les continents africains et asiatiques.

Selon différents rapports de l'OMS, près de deux tiers des causes de déficiences visuelles pourraient être évitées, tant par la prévention que par les traitements [World Health Organization, 2010a]. Ce constat a donné lieu au lancement en 1999 d'un programme mondial baptisé « VISION 2020 : The Right to Sight », initié par l'OMS et l'IAPB[2], une organisation regroupant près de vingt ONGs, ainsi que plusieurs associations professionnelles, instituts et corporations dans le domaine de la santé et de l'ophtalmologie [IAPB, 2010]. Par des campagnes de financement, d'information et de coordination, ce projet vise à développer les infrastructures et les technologies, à assurer la formation d'ophtalmologues, ainsi qu'à sensibiliser l'opinion publique et les dirigeants afin que chaque pays puisse mettre en place des politiques sanitaires adaptées [Foster and Resnikoff, 2005; World Health Organization, 2005]. Initié il y a plus de 10 ans, ce programme a déjà montré des résultats encourageants, et les prévisions pour 2020 indiquent un impact sensible sur le nombre de déficients visuels dans les pays émergents [Frick and Foster, 2003; IAPB, 2010].

Toutefois, même en améliorant la qualité des soins et de la prise en charge dans les pays en voie de développement, le nombre de déficients visuels ne devrait cesser d'augmenter à l'échelle mondiale. Le nombre d'aveugles continue d'ailleurs de croître de près de 2 millions chaque année malgré les différents programmes tels que VISION 2020. Ces chiffres s'expliquent par l'accroissement de l'espérance de vie. Une grande partie des maladies cécitantes étant liée à l'âge, le vieillissement de la population entraîne par conséquent une augmentation du nombre de personnes touchées.

[1] Handicaps - Incapacités – Dépendance
[2] International Agency for the Prevention of Blindness

De nombreuses études épidémiologiques ont en effet montré que la prévalence de la déficience visuelle augmente nettement avec l'âge (elle peut être multipliée par 20 entre 50 et 80 ans). Ainsi, en Angleterre, plusieurs enquêtes ont montré qu'environ 20 % de la population âgée de plus de 75 ans a une acuité visuelle inférieure à 6/12 [Wormald et al., 1992]. Des résultats similaires ont été trouvés aux Etats Unis, dans les enquêtes de Salisbury [West et al., 1997], de Baltimore [Tielsch et al., 1990] et de plusieurs autres villes américaines [Salive et al., 1992]. En Australie, deux grandes études menées dans les années 90 (Melbourne Visual Impairment Project et Blue Mountains Eye Study) ont également montré une augmentation constante des troubles visuels avec l'âge, touchant moins d'un pourcent de la population à 50 ans, puis entre 30 et 40 % au-delà de 85 ans [Attebo et al., 1996; Taylor et al., 2005; VanNewkirk et al., 2001]. Les prévalences des troubles visuels chez les personnes âgées observées dans six différentes études conduites dans des pays industrialisés sont compilées dans la Figure 1, tirée de [Klaver CW et al., 1998], qui illustre clairement cette relation entre vieillissement et malvoyance.

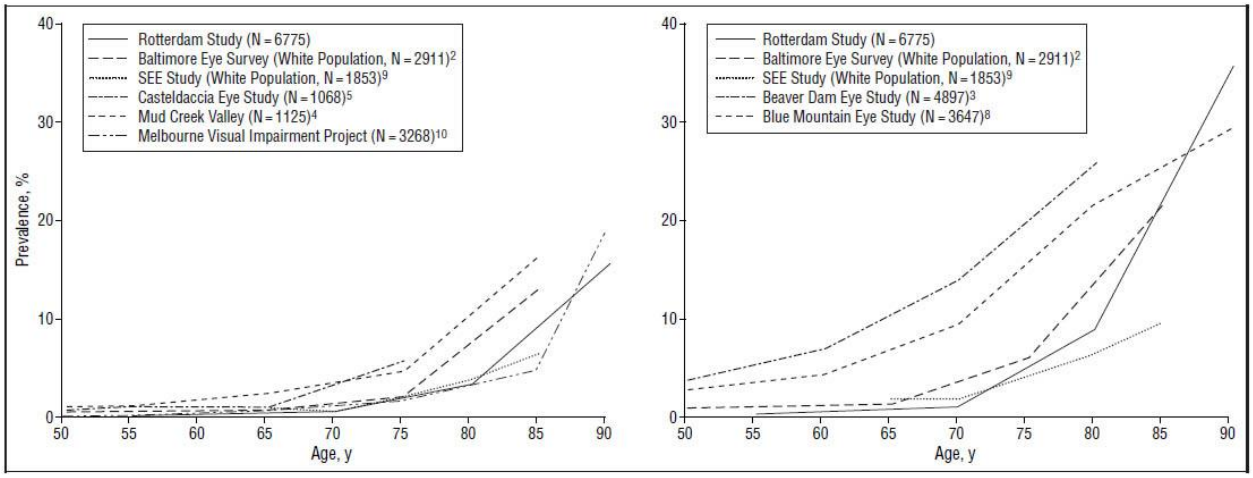


Figure 1 Prévalence de la basse-vision dans la population en fonction de l'âge, selon le critère établit par l'OMS à gauche (acuité visuelle inférieure à 20/60), ou celui employé aux Etats-Unis à droite (inférieure à 20/40).

D'autre part, si les améliorations dans le traitement de maladies comme les glaucomes ou le diabète pourront réduire le nombre de personnes souffrant de pertes de vision imputables à ces pathologies, beaucoup d'autres ne sont pas traitables, telles que la DMLA[1], qui reste la principale cause de cécité dans les pays occidentaux [Margrain, 2000]. Le nombre de malvoyants devrait donc continuer d'augmenter, allant même jusqu'à doubler d'ici 2030 selon certaines projections [Foran et al., 2000; Taylor et al., 2005]. Les prévisions sur l'évolution du nombre d'aveugles profonds sont tout aussi alarmantes, puisqu'un un

[1] Dégénérescence maculaire liée à l'âge.

rapport de 1995 prévoyait qu'il passe de 22 à 54 millions d'ici à 2020 [Thylefors et al., 1995], chiffre revu à la hausse au début des années 2000 par Kevin Frick qui, en tenant compte de nouvelles projections démographiques, estimait qu'il pourrait s'élever à 76 millions [Frick and Foster, 2003].

Parallèlement aux campagnes de prévention, à l'amélioration de la prise en charge, et aux recherches cliniques tentant de développer de nouveaux traitements, il est par conséquent crucial de proposer des solutions permettant d'améliorer la qualité de vie des non-voyants, car leur nombre ne devrait cesser d'augmenter. Ceci constitue l'objet de cette thèse. Afin de mettre en œuvre des systèmes de suppléance adaptés, il est nécessaire de cerner au mieux les besoins et les attentes des déficients visuels. Nous proposerons donc dans le premier chapitre un tour d'horizon du handicap visuel en présentant les différents types de déficiences visuelles, leurs causes, ainsi que leurs conséquences.

Nous présenterons également un état de l'art des systèmes visant à compenser la perte ou l'absence de vision. D'une part les dispositifs basés sur une approche fonctionnelle, répondant à des besoins spécifiques, et d'autre part les systèmes génériques, ou holistiques, qui se regroupent en deux catégories : les systèmes de substitution sensorielle, restituant l'information visuelle par l'intermédiaire d'une autre modalité sensorielle (généralement l'audition ou le toucher), et les neuroprothèses visuelles. Ces dernières reproduisent à la surface du relais visuel implanté les images acquises par une caméra, en respectant leur configuration spatiale, un pixel correspondant à une électrode. Malheureusement, les neuroprothèses actuelles souffrent encore d'une perte de résolution spatiale trop importante, puisqu'une image ne sera restituée que par une matrice de quelques dizaines de points, rendant ces systèmes inadaptés à une utilisation au quotidien. Ces limitations sont de même nature dans le cas des dispositifs de substitution sensorielle : la quantité d'informations visuelles nécessaire à l'interprétation d'une scène naturelle est bien trop importante par rapport à la résolution de l'interface de restitution (auditive, tactile, ou par micro-stimulation). Ces systèmes se montrent par conséquent inefficaces dans des environnements visuels complexes, et ils ne sont donc qu'extrêmement peu utilisés en dehors des laboratoires de recherches.

Comment alors utiliser les images acquises par des caméras embarquées pour fournir des informations exploitables par un non-voyant, et ce malgré les faibles résolutions de sortie ? Pour répondre à cette problématique, nous proposerons dans cette thèse une approche alternative, consistant à intégrer des méthodes de vision artificielle, afin de prétraiter la scène visuelle, et de ne restituer au non-voyant que les informations extraites pertinentes. Nous détaillerons dans le deuxième chapitre la mise en place et à l'évaluation d'un nouveau système d'assistance aux déficients visuels reposant sur ce principe, baptisé Navig.

Nous montrerons notamment que grâce à la reconnaissance de formes en temps réel et à la synthèse de sons spatialisés, ce système permet de restaurer des boucles visuomotrices qui rendent à nouveau possibles certaines fonctions visuelles comme la localisation et la préhension d'objets. La navigation étant une autre tâche critique pour les non-voyants, nous avons également incorporé au dispositif des fonctions de guidage basées sur le positionnement par satellites et sur un système d'information géographique adapté. La trop faible précision de localisation du GPS nous a amenés à développer une nouvelle méthode de positionnement hybride, combinant les données satellites et inertielles à la reconnaissance de cibles visuelles géolocalisées. L'utilisation de la vision artificielle a ainsi permis d'améliorer les performances de localisation et d'obtenir une erreur moyenne généralement inférieure à 5 mètres, rendant possible le guidage et la navigation en temps-réel d'un piéton non-voyant.

Dans le troisième chapitre, nous nous concentrerons sur le moteur de reconnaissance de formes, qui constitue le cœur du dispositif Navig. L'algorithme que nous avons utilisé, Spikenet, s'inspire du fonctionnement du système visuel humain [Delorme and Thorpe, 2003; Thorpe et al., 2004]. Il repose notamment sur un codage de l'information visuelle par latence, et des représentations sous forme d'arêtes orientées, telles que celles observées dans le cortex visuel primaire. L'emploi de Spikenet au sein du système Navig nous a permis de mettre à jour certaines de ses limites, telles que sa tolérance aux transformations affines, ou le temps requis pour rechercher de nombreux objets simultanément. Afin d'augmenter sa vitesse de traitement, et d'enrichir l'information extraite des motifs à apprendre, nous avons développé un nouvel algorithme de reconnaissance de formes multi-résolutions reposant sur une détection en cascade qui combine plusieurs traitements successifs à différentes échelles. Par rapport à l'algorithme originel mono-échelle, cette architecture permet de capturer un spectre de fréquences spatiales plus large. Les traitements à faible résolution permettent ainsi d'améliorer la tolérance aux déformations de l'image, alors que les hautes fréquences spatiales, plus discriminantes, maintiennent une précision suffisamment élevée. De par son fonctionnement en plusieurs passes successives, cette nouvelle architecture permet de plus de diminuer les temps de traitement grâce à une première couche rapide, filtrant les objets à rechercher et les régions d'intérêt dans la phase suivante à haute résolution, plus coûteuse en temps de calcul.

Les différentes contributions de cette thèse s'articulent donc autour de la vision artificielle, ou vision par ordinateur, que l'on définit comme l'ensemble des méthodes visant à extraire de façon automatique des informations haut-niveau à partir d'images ou de vidéos. Comme nous venons de l'évoquer, nous présenterons notamment la mise en place d'un système d'aide aux non-voyants (et malvoyants profonds) basé sur l'utilisation de caméras embarquées et d'algorithmes de reconnaissance de formes, développé à partir

d'une analyse préalable des besoins de la population souffrant de déficiences visuelles. Nous nous intéresserons en particulier aux aspects du système relatifs à la vision, à savoir la boucle d'interaction permettant à l'utilisateur de localiser des objets d'intérêt, ainsi que la méthode de positionnement que nous avons développé, utilisant la reconnaissance de points de repère visuels. Nous détaillerons également l'architecture d'un nouvel algorithme de reconnaissance de formes bio-inspiré, développé au cours de cette thèse, pouvant être appliqué dans le cadre de l'aide aux non-voyants mais aussi dans tout autre contexte nécessitant de reconnaitre et localiser des cibles visuelles en un minimum de temps. Pour terminer, nous proposerons dans la conclusion générale une synthèse des résultats obtenus et discuterons des perspectives pouvant faire suite à ces travaux.

I. Contexte et état de l'art

Sommaire de section

1. Le handicap visuel

1.1 Classification des déficiences visuelles

La vision est un processus psychosensoriel complexe, résultant de l'interaction de nombreux facteurs. Ainsi, les rayons lumineux, lorsqu'ils traversent la cornée puis le cristallin, sont concentrés pour former une image nette sur la rétine, à l'arrière de l'œil, où les photorécepteurs, convertissent ce signal en messages électriques. Ces signaux sont ensuite acheminés jusqu'au cortex visuel primaire par le biais de plusieurs relais synaptiques, notamment grâce au nerf optique, composé des terminaisons nerveuses des neurones ganglionnaires de la rétine, qui se projettent dans le thalamus. La sensation visuelle est le résultat de différents traitements effectués dans le cortex visuel et les aires associatives, permettant la perception de l'environnement par l'appréciation des formes, des couleurs, du mouvement, des distances… Selon l'origine et l'importance des affections du système visuel, les conséquences perceptives et fonctionnelles pourront toucher des composantes très différentes de la vision. Cependant, seulement deux aspects sont généralement considérés dans l'évaluation de la déficience visuelle : l'acuité et le champ visuel. L'acuité visuelle mesure le sens morphoscopique, c'est-à-dire la capacité de l'œil à distinguer les détails de l'espace, alors que l'examen du champ visuel évalue la portion de l'espace perçue en regardant face à soi.

L'OMS, dans la classification internationale des maladies (CIM-10) et dans la classification internationale du fonctionnement, du handicap et de la santé (CIF), définit cinq catégories de déficiences qui tiennent compte à la fois de la baisse de l'acuité visuelle et de la réduction du champ visuel. Les catégories 1 et 2 correspondent à ce que l'on nomme communément la malvoyance (également appelée basse vision ou vision réduite), et celles de 3 à 5 à la cécité. Les critères d'évaluation reposent toujours sur une baisse d'acuité visuelle ou sur une diminution du champ visuel.

Le champ visuel s'exprime en degrés. La norme étant 180°, on parle de malvoyance lorsqu'il est inférieur à 20° et de cécité en dessous de 10°. L'acuité visuelle, qui mesure le pouvoir séparateur de l'œil à une distance donnée (c'est-à dire la faculté à discerner deux points distincts), est généralement notée sous la forme d'une fraction. Le numérateur correspond à la distance à laquelle se trouve l'optotype[1], et le dénominateur à la distance

[1] Un optotype est un tableau contenant une échelle visuelle constituée de figures ou de caractères. Parmi les plus courants on peut citer les tests de Snellen, les anneaux de Landolt ou encore les échelles Monoyer et Parinaud.

maximale à laquelle un individu à la vision normale (10/10) peut distinguer le même motif. Par exemple, une acuité de 1/20ème signifie qu'un objet perçu à 20 mètres par un individu ayant une vision normale doit être placé à 1 mètre de la personne déficiente visuelle pour être perçu de la même façon.

Selon les systèmes d'unité de mesures, les distances sont exprimées en mètres ou en pieds et les conventions varient : dans les pays anglo-saxons, la référence est 20/20 et le numérateur toujours constant, alors que dans le système métrique elle est rapportée en dixièmes. Cependant, il suffit de les convertir sous forme décimale pour obtenir un moyen de comparaison. Une acuité visuelle de 1/20 (selon le système français) est par exemple équivalente à 20/400 (selon la norme anglo-saxonne), ou à 0,10 si exprimée selon les recommandations européennes (EN ISO 8596).

Catégorie OMS	*Acuité visuelle et champ de vision*	*Type d'atteinte visuelle*	*Type de déficience visuelle*
Catégorie 1	*1/10 < AVC < 3/10 champ visuel d'au moins 20°*	*Basse vision*	*Moyenne*
Catégorie 2	*1/20 < AVC < 1/10*	*Basse vision*	*Sévère*
Catégorie 3	*1/50 < AVC < 1/20 5° < champ visuel < 10°*	*Cécité*	*Profonde*
Catégorie 4	*AVC < 1/50 mais perception lumineuse préservée champ visuel < 5°*	*Cécité*	*Presque totale*
Catégorie 5	*Cécité absolue, absence de perception lumineuse*	*Cécité*	*Totale*

Tableau I-1 Classifications des déficiences visuelles selon l'OMS (AVC signifie Acuité Visuelle Corrigée).

Les critères retenus par la dernière révision de la classification internationale des maladies de l'OMS sont fournis dans le Tableau I-1. Si ces définitions constituent désormais la référence mondiale dans l'épidémiologie des déficiences visuelles, la plupart des états ont malgré tout conservé leur propre définition du handicap visuel. Ainsi en France, dans l'attribution du taux d'incapacité[1], la cécité complète correspond à une vision totalement abolie, sans aucune perception de lumière. Elle correspond à la cécité totale de l'OMS (catégorie 5). La quasi-cécité et la cécité professionnelle sont définies par une vision strictement inférieure à 1/20 pour le meilleur œil, ou par un champ visuel réduit à 20 degrés

[1] Le taux d'incapacité est un indice fixé par la Caisse Primaire d'Assurance Maladie (CPAM) calculant le niveau de dépendance de la personne.

(correspondant à la catégorie 3 de l'OMS, dite cécité partielle). Dans la plupart des pays anglo-saxons, les seuils sont plus larges. On parle en effet de cécité légale lorsque l'acuité visuelle du meilleur œil est inférieure ou égale à 1/10 et de malvoyance lorsque l'acuité visuelle est inférieure à 5/10. Cette définition de la cécité inclut donc la baisse de vision (catégorie 2) de l'OMS. C'est le cas pour les Etats-Unis, le Canada ainsi que pour plusieurs pays européens (Pays-Bas, Royaume-Uni, pays scandinaves). A l'inverse, les critères en vigueur en Allemagne sont plus sévères, car seules les personnes ayant une acuité visuelle inférieure à 1/50 pour leur meilleur œil sont légalement considérées comme aveugles.

En suivant la classification de l'OMS, parmi les 1 700 000 personnes qu'on estime atteintes de déficits visuels en France, près de 560 000 personnes seraient des malvoyants légers avec une acuité visuelle ou un champ visuel périphérique étroit mais sans incapacité visuelle sévère déclarée en vision centrale. Environ 932 000 individus seraient malvoyants moyens (catégorie 1), 146 000 malvoyants profonds (catégorie 2), et enfin 61 000 non-voyants (catégories 3, 4 et 5). La prévalence de ces catégories par tranche d'âge est reportée dans la Figure I-1, d'après les résultats des enquêtes HID (Handicaps, incapacités, dépendance) de 1998 et Domicile Ordinaire de 1999-2000 [Observatoire Régional de la Santé des Pays de la Loire, 2000].

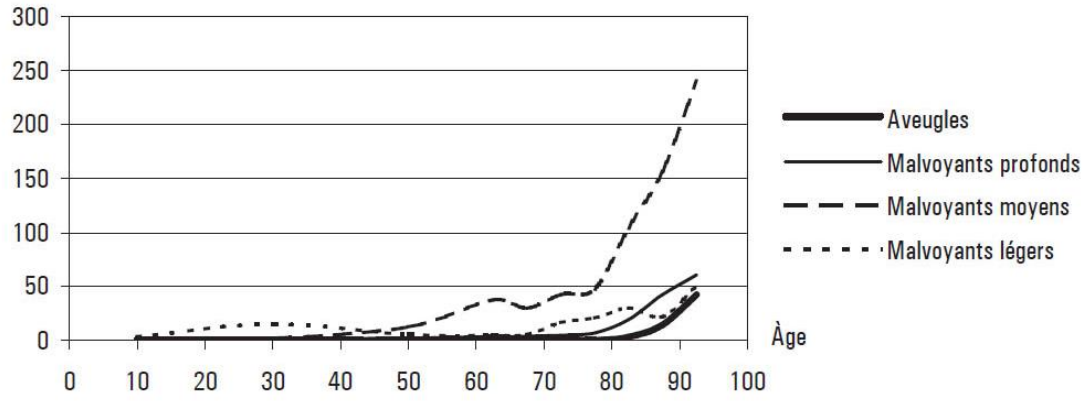


Figure I-1 Prévalence de la déficience visuelle en France métropolitaine selon l'âge et le degré de sévérité (exprimés en taux pour 1000).

1.2 Causes de déficiences visuelles

Selon les dernières estimations de l'OMS, la première cause de cécité à l'échelle mondiale serait la cataracte (39 %), suivie par les troubles de la réfraction non corrigés (18 %) et les différentes formes de glaucomes (10 %). Viennent ensuite les dégénérescences maculaires liées à l'âge, ou DMLA (7 %), l'opacité cornéenne (4,3 %), la rétinopathie diabétique (4 %), les trachomes (3 %), diverses maladies de l'œil chez l'enfant (3 %), et enfin l'onchocercose, qui ne représente plus que 0,7 % des cas compte tenu des résultats de la lutte contre la maladie entreprise par l'OMS en Afrique occidentale depuis vingt ans [IAPB, 2010; Thylefors et al., 1995].

Concernant la malvoyance, ce sont les troubles de la réfraction non corrigés qui arrivent en tête des causes de déficience visuelle (43 % des cas). Ces amétropies (myopie, presbytie, hypermétropie et astigmatisme) ont longtemps été négligées, car la plupart des estimations épidémiologiques prenaient en compte l'acuité visuelle corrigée. Pourtant à l'échelle mondiale, y compris dans les pays industrialisés, un nombre important de ces pathologies reste non corrigées en raison des coûts trop importants des aides visuelles ou de l'absence de dépistage [Resnikoff et al., 2008]. Après les troubles de la réfraction, la cataracte constitue le deuxième facteur de malvoyance (33 %). Les trachomes, DMLA, rétinopathies diabétiques et opacités cornéennes représentent environ 1 % des cas pour chacune, et les glaucomes 2 %. Une grande proportion des causes (près de 18 %) reste indéterminée faute de données épidémiologiques détaillées [Pascolini and Mariotti, 2012; Resnikoff et al., 2004]. La répartition des différentes causes de cécité et de malvoyance est compilée dans la Figure I-2.

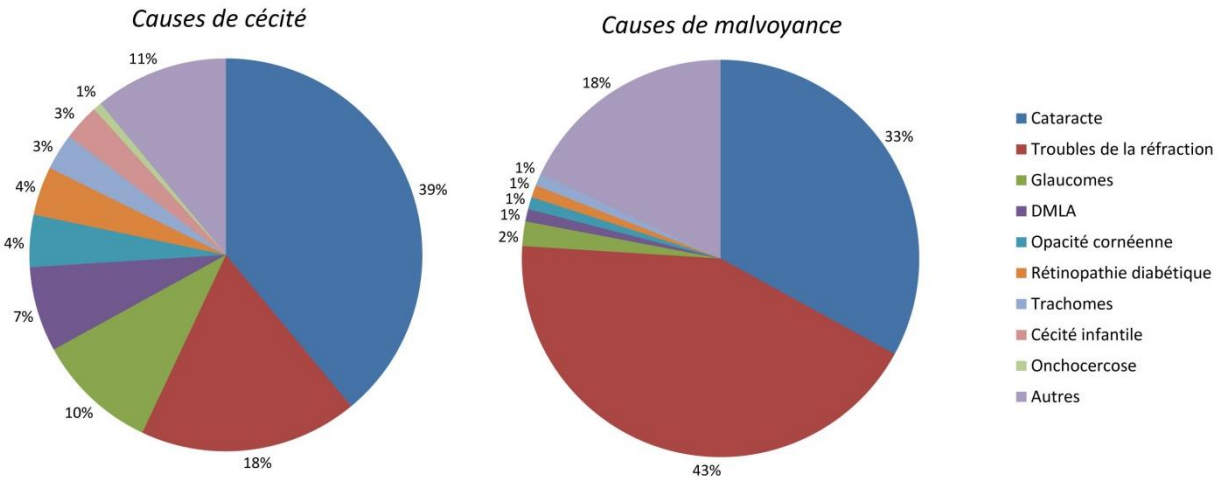


Figure I-2 Principales causes de malvoyance et de cécité à l'échelle mondiale.

1.2.1 Description des pathologies

Une description succincte des principaux troubles et pathologies responsables de malvoyance ou de cécité est proposée dans cette section. Les structures oculaires affectées par chacune d'elles sont résumées dans la Figure I-3, extraite de [Congdon NG et al., 2003].

Causes of Visual Impairment	Affected Ocular Structures
Refractive Error (Myopia, Hyperopia)	Lens, Sclera, Retina
Age-Related Causes	
Cataract	Lens
Glaucoma (Angle-Closure)	Anterior Chamber, Optic Nerve
Glaucoma (Open Angle)	Trabecular Meshwork, Sclera (in Children), Optic Nerve
Age-Related Macular Degeneration	Retina, Macula, Choroid
Infectious Causes	
Trachoma	Lids and Lashes, Cornea
Onchocerciasis	Cornea, Trabecular Meshwork, Ciliary Body, Iris, Retina, Macula, Choroid, Optic Nerve
HIV and HIV-Associated Infection	Lacrimal Gland, Cornea, Choroid, Retina, Optic Nerve
Nutritional and Metabolic Causes	
Vitamin A Deficiency	Conjunctiva, Cornea, Retina
Diabetes	Trabecular Meshwork, Lens, Vitreous Body, Retina, Macula, Optic Nerve
Trauma	All

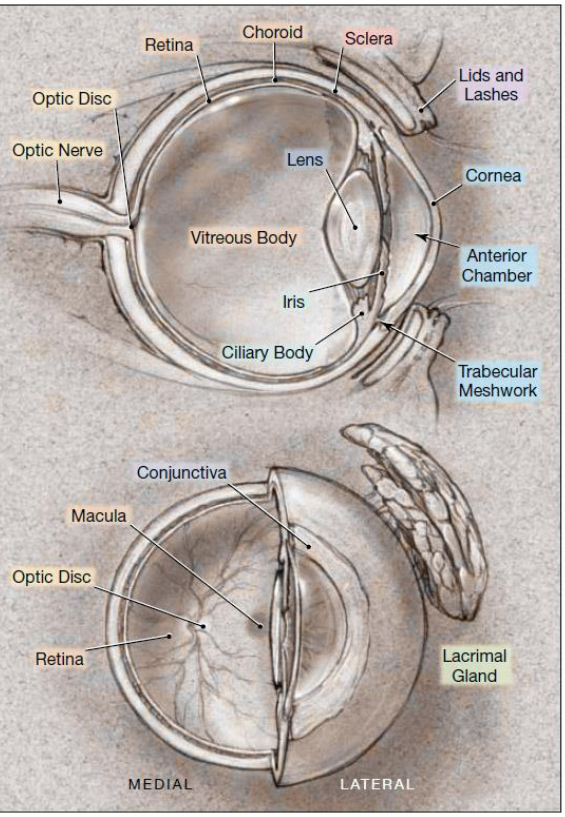


Figure I-3 Structures oculaires affectées par chacune des causes de pathologies visuelles les plus fréquentes.

Cataracte

La cataracte, première cause de cécité et de malvoyance dans le monde, correspond à une opacification du cristallin qui entraîne une baisse graduelle de l'acuité visuelle, jusqu'à la cécité si celle-ci n'est pas traitée. Bien que certains enfants puissent naître avec cette maladie, elle se développe généralement avec le vieillissement (plus d'une personne sur cinq à partir de 65 ans est touchée, plus d'une sur trois à partir de 75 ans et près de deux sur trois après 85 ans). Le traitement est chirurgical (extraction du cristallin et implantation d'une lentille intraoculaire) et résulte dans la grande majorité des cas en une réhabilitation visuelle immédiate et complète. Ces interventions sont très courantes dans les pays

développés (plus d'un million d'opérations sont réalisées chaque année aux États-Unis, et en France il s'agit de l'acte chirurgical le plus pratiqué), mais encore trop rares dans les pays en voie de développement par manque de prise en charge.

Glaucome

Les glaucomes sont une famille de pathologies qui se caractérise par des dommages au nerf optique et une surpression intraoculaire, entraînant une diminution progressive et irréversible du champ visuel. La prévention du glaucome, dont le développement est insidieux et peu douloureux, nécessiterait un dépistage systématique pour être efficace. S'il n'est pas possible de recouvrer le champ visuel perdu, la chirurgie laser permet de stopper la progression de la maladie pour certains types de glaucomes diagnostiqués suffisamment tôt.

Dégénérescences maculaires liées à l'âge

La dégénérescence maculaire liée à l'âge (DMLA) est la première cause de cécité dans les pays industrialisés. Elle correspond à une atrophie de l'épithélium pigmentaire rétinien de la fovéa, zone centrale de la rétine, qui se traduit par une altération de la vision centrale. Elle entraîne d'abord des gêne à la lecture, à la reconnaissance des visages ou à la conduite, et peut évoluer vers des déficiences visuelles plus graves allant jusqu'à la cécité. Sa prévalence augmente avec l'âge. Ainsi, elle représente la première cause de cécité chez les personnes de plus de 50 ans et touche 25 % des personnes à partir de 80 ans. Il n'existe actuellement pas de traitement (curatif ou préventif). Avec le vieillissement de la population, le problème de santé publique que constitue la DMLA risque donc de s'accentuer.

Rétinopathie diabétique

La rétinopathie diabétique est la conséquence de lésions des vaisseaux capillaires de la rétine dues au diabète. Dans sa forme grave, proliférante, un œdème maculaire peut se développer avec l'apparition de néo-vaisseaux, entraînant une réduction considérable de la vision. Sa prévention passe par un bon contrôle du diabète et un suivi ophtalmologique régulier. Une fois la maladie déclarée, le traitement repose sur la photo-coagulation laser des vaisseaux capillaires de la rétine afin de prévenir la perte fonctionnelle. Environ 10 % des patients développent une déficience visuelle grave après 10 ans de diabète car même dans les pays développés le recours aux soins est freiné par le manque de sensibilisation du grand public.

Trachome

Le trachome est une maladie infectieuse qui touche environ 84 millions de personnes dont 8 millions ont une déficience visuelle. Il est provoqué par un parasite, *Chlamydia trachomatis*, entraînant des conjonctivites chroniques avec complications palpébrales et cornéennes aboutissant à la cécité. Actuellement responsable d'environ 3% des cas de cécité dans le monde, il était par le passé endémique dans la plupart des pays. Sa prévention passe par une amélioration de l'hygiène et le nombre de trachomateux tend donc à diminuer grâce au développement socio-économique et aux programmes de lutte contre cette maladie. Néanmoins le trachome continue à être fréquent dans certains pays en voie de développement d'Afrique, d'Asie, d'Amérique du Sud, (espace en trop) et du Moyen-Orient.

Onchocercose

L'onchocercose, couramment appelée "cécité des rivières", est une maladie provoquée par un vers nommé *Onchocerca volvulus* qu'on trouve dans certaines rivières d'Afrique et d'Amérique centrale. Chez les personnes exposées, on retrouve ces parasites dans tous les tissus oculaires, excepté le cristallin, où elles provoquent une inflammation, des hémorragies et d'autres complications qui conduisent finalement à la cécité. Malgré les progrès accomplis dans la lutte contre la maladie (notamment par l'éradication des simulies, les moucherons vecteurs du parasite), on estime qu'il y a un demi-million de personnes aveugles en raison de la cécité des rivières. Il existe maintenant un traitement consistant en une dose annuelle d'antiparasitaire (l'Ivermectine).

Opacité cornéenne

Les déficiences visuelles d'origine cornéenne comprennent un ensemble de pathologies infectieuses, inflammatoires ou traumatiques, à l'origine de cicatrices cornéennes plus ou moins opaques gênant la vision. Parmi les causes importantes de cécité cornéenne on compte les trachomes, les traumatismes oculaires, les ulcérations de la cornée, la xérophtalmie ou encore la conjonctivite gonococcique, la lèpre et l'onchocercose. Le seul traitement curatif actuellement disponible est la greffe de cornée. Mais l'accès à cette chirurgie est très difficile y compris dans les pays développés, du fait du manque de donneurs.

Cécité infantile

Les carences en vitamine A restent à ce jour la première cause de cécité infantile. Peu fréquentes dans les pays développés, elles sont communes dans de nombreuses régions d'Afrique, où jusqu'à 500 000 enfants victimes de malnutrition perdent la vue chaque année

à cause de ces carences, dont près de la moitié décèdent dans l'année suivant l'apparition de la cécité. Les premiers signes d'une carence en vitamine A sont l'héméralopie (déficit de l'adaptation à l'obscurité) et la cécité nocturne. Une carence sévère ou prolongée entraîne souvent des xérophtalmies et kératoconjonctivites, c'est-à-dire des lésions de la cornée souvent associées à des ulcères, à l'origine d'une cécité irréversible. Les autres causes de cécité infantile, bien que moins courantes, sont la rougeole, la conjonctivite néonatale, la microphtalmie, les cataractes congénitales (qui peuvent résulter de la contraction de la rubéole durant la grossesse), et enfin certaines maladies génétiques comme les rétinites pigmentaires.

1.2.2 Epidémiologie

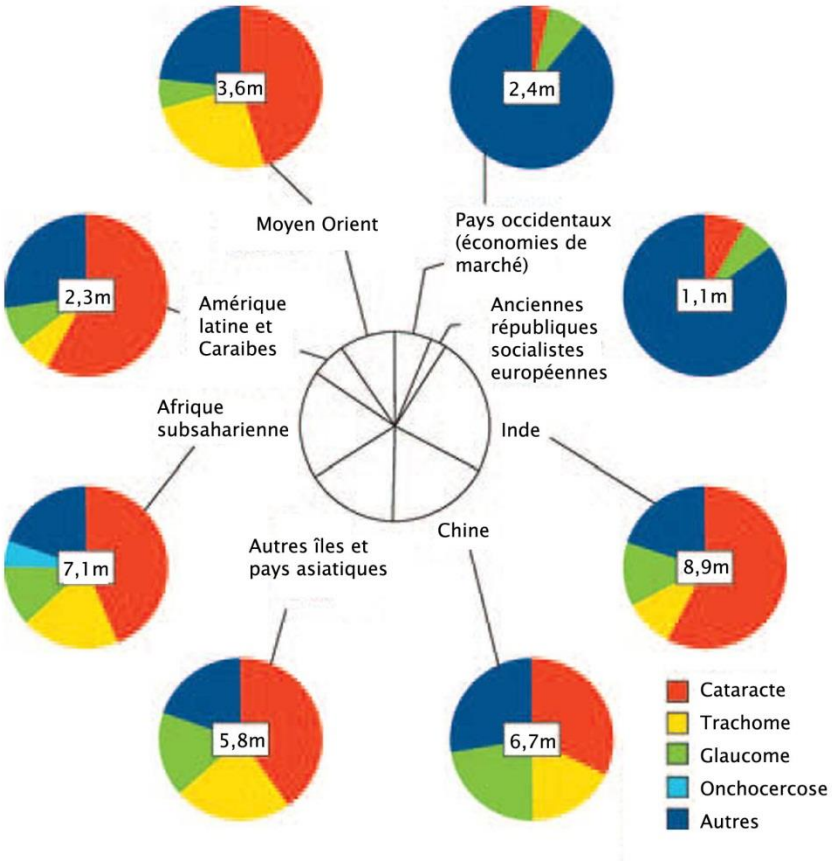


Figure I-4 Causes de cécité dans les différentes régions définies par l'OMS (figure adaptée de [Thylefors et al., 1995]).

Les principales causes de cécité et de malvoyance varient considérablement d'une région à l'autre, car elles sont en grande partie déterminées par le développement socio-économique et la disponibilité des soins de santé. Ainsi certaines pathologies comme les trachomes, onchocercoses ou déficits en vitamine A ne touchent presque exclusivement que

les pays du Sud, d'autres comme la cataracte ou les troubles de la réfraction sont présents dans toutes les populations, mais leur prise en charge étant bien meilleure dans les pays industrialisés, le nombre et le type de déficiences résultantes diffèrent grandement [Frick and Foster, 2003]. Ceci explique que 75% des malvoyants à l'échelle mondiale vivent dans des pays en voie de développement, et que la prévalence de la cécité soit, dans ces pays près de deux fois supérieure à celle observée en Europe [World Health Organization, 2005]. La répartition des différentes causes de cécité dans chacune des régions définies par l'OMS, donnée par la Figure I-4 illustre bien ces inégalités.

Il existe des moyens de prévention et des traitements peu coûteux pour la plupart des pathologies responsables de troubles visuels. On estime que jusqu'à 80% de celles-ci pourraient être évitées [Foster and Resnikoff, 2005; Pascolini and Mariotti, 2012]. Avec des ressources suffisantes et des programmes de santé comme VISION 2020, de nombreuses maladies infectieuses responsables de troubles visuels pourraient être éradiquées, et d'autres pathologies comme la cataracte, opérées de façon plus systématique. La prévalence et la répartition des déficiences visuelles dans les pays émergents devraient alors se rapprocher de celles observées actuellement dans les pays développés, où les trois principales causes de malvoyance et de cécité sont la DMLA, la rétinopathie diabétique et les glaucomes [Observatoire Régional de la Santé des Pays de la Loire, 2000].

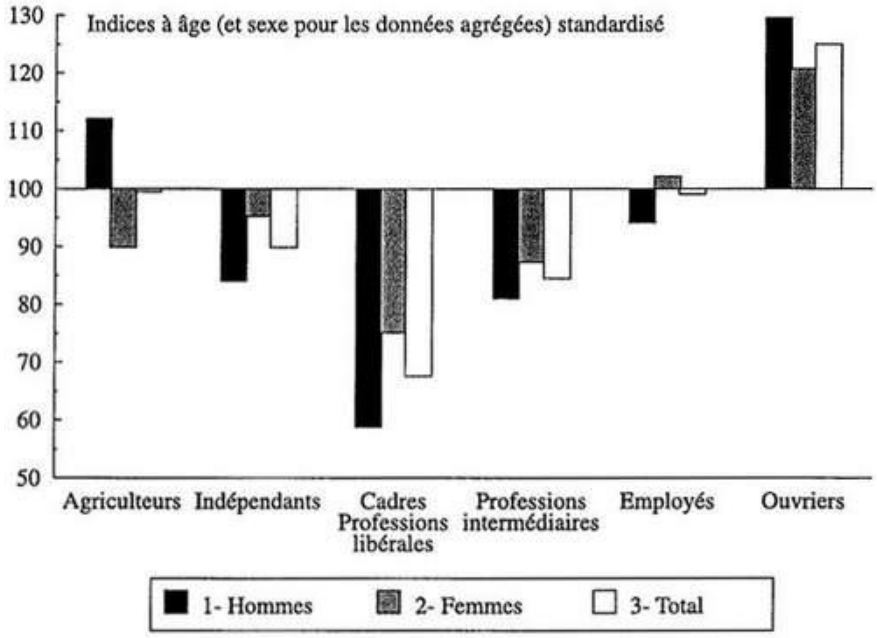


Figure I-5 Variabilité du nombre de déficiences en France métropolitaine selon le sexe et la catégorie socioprofessionnelle (figure tirée [Mormiche and Boissonnat, 2003], à partir de données de l'enquête HID).

Ces inégalités face au handicap visuel se retrouvent non seulement parmi les différentes régions du monde, mais également au sein d'un même pays en fonction de la catégorie socioprofessionnelle et du sexe [Mormiche and Boissonnat, 2003; Sander et al., 2005], tel qu'illustré dans la Figure I-5. Elles varient également selon le groupe ethnique [The Eye Diseases Prevalence Research Group, 2004], mais le facteur ayant le plus d'incidence est l'âge [World Health Organization, 2010b]. Les personnes de plus de 50 ans représentent en effet respectivement 65 et 82 % des malvoyants et aveugles à l'échelle mondiale [Klaver CW et al., 1998].

1.3 Handicap et autonomie

Le handicap est une notion complexe comprenant de nombreux aspects, touchant aussi bien à la santé qu'à la situation sociale. Afin de clarifier ce concept, l'OMS, en se basant sur les travaux de Philip Wood, a proposé en 1980 une classification internationale des handicaps, ou CIH [Organisation Mondiale de la Santé, 1980], qui distingue trois composantes du handicap : les déficiences, les incapacités engendrées par une déficience, et enfin les désavantages qui en résultent pour la personne, décrits dans la Figure I-6. Ces trois niveaux, même s'ils ne s'inscrivent pas toujours dans un enchaînement linéaire, permettent d'appréhender la dynamique d'un processus qui lie dimension biomédicale et dimension sociale.

Dans cette classification, les déficiences correspondent à l'aspect lésionnel du handicap, et sont définies comme les altérations d'une structure ou fonction psychologique, physiologique ou anatomique. Les incapacités se rapportent quant à elles aux aspects fonctionnels et correspondent à une réduction partielle ou totale de la capacité à accomplir une activité résultant d'une déficience. Elles sont classées en différentes catégories : le comportement, la communication, les soins corporels, la locomotion, l'utilisation du corps dans certaines tâches, ou encore les manipulations. Enfin, les désavantages désignent les préjudices résultant d'une déficience ou d'une incapacité qui limitent ou interdisent l'accomplissement d'un rôle social normal (en rapport avec l'âge, le sexe, les facteurs socioculturels). Ils correspondent donc à l'aspect situationnel du handicap. Parmi les principaux types de désavantages, on relèvera ceux touchant l'orientation, l'indépendance physique, la mobilité, la scolarité, les activités occupationnelles, l'intégration sociale, et l'indépendance économique.

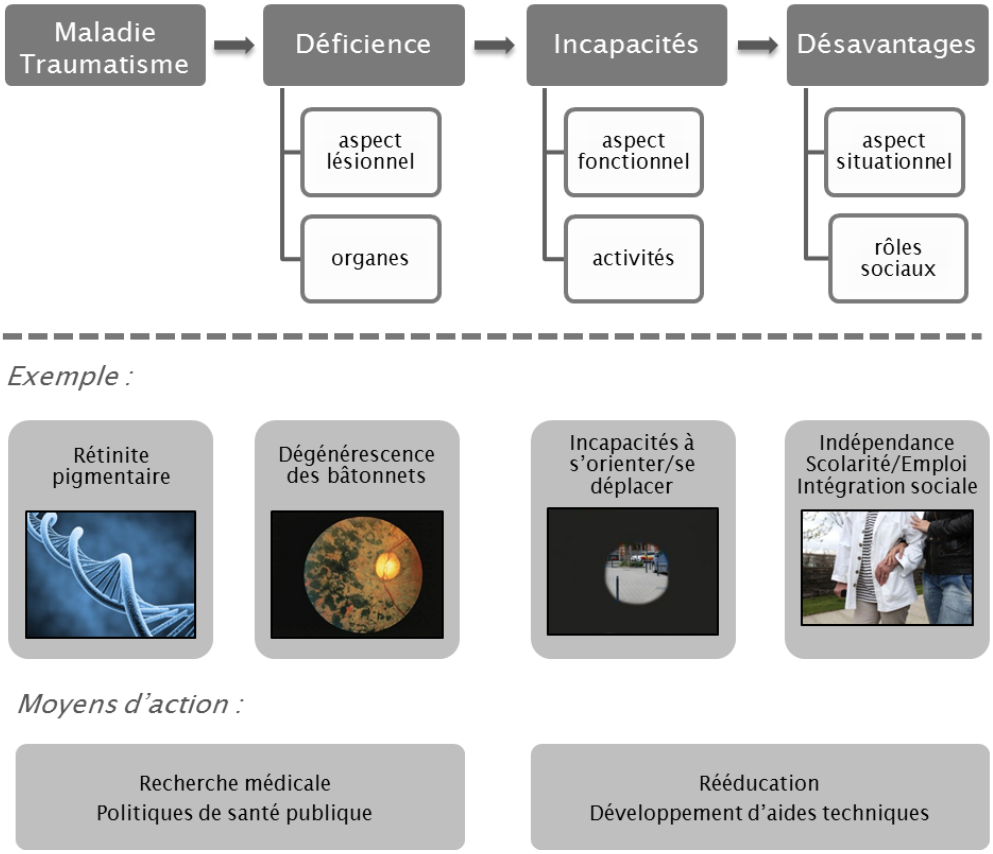


Figure I-6 Schéma de Wood sur les différentes composantes du handicap, reprises dans la classification internationale du Handicap de l'OMS.

En 2001, la Classification Internationale du Fonctionnement, du Handicap et de la Santé (CIF) a permis d'introduire différentes révisions, et se substitue depuis à la CIH de 1980 [Organisation Mondiale de la Santé, 2001]. Reflétant l'évolution des normes internationales et des représentations sociales du handicap, la CIF bascule d'un modèle médical (excluant par exemple les dimensions environnementales et personnelles), à des modèles fonctionnels et sociaux. Contrairement à la linéarité reprochée à la CIH, les concepts introduits dans la CIF permettent de représenter la pluralité des interférences entre plusieurs composantes (illustrés dans la Figure I-7) :

- les activités que font les individus et les domaines de la vie auxquels ils participent,
- les facteurs environnementaux qui influencent leur participation,
- les fonctions organiques et les structures anatomiques des individus,
- les facteurs personnels.

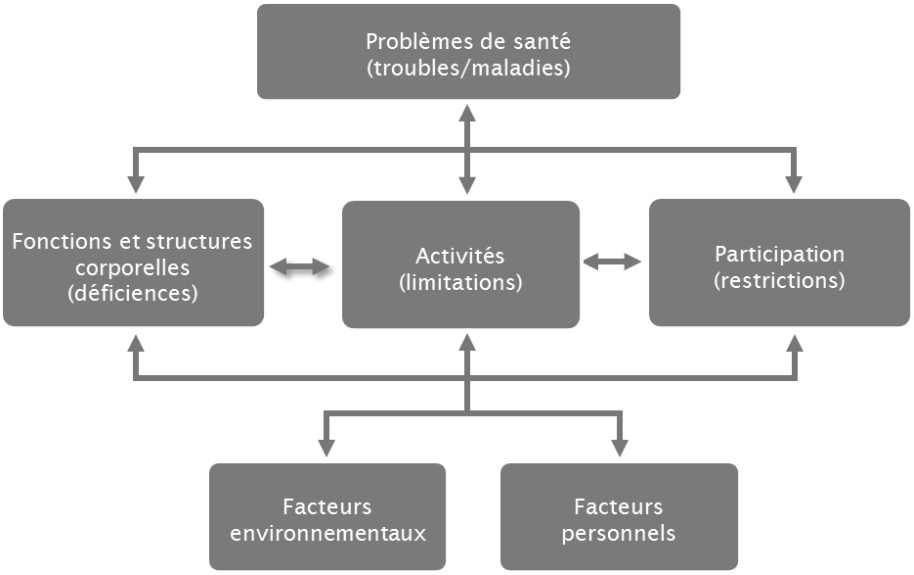


Figure I-7 Schéma conceptuel de la CIF

Les incapacités et désavantages de la CIH sont désormais définis dans cette nouvelle terminologie comme des limitations d'activités (difficultés que rencontre une personne dans l'exécution d'activités), et des restrictions de participation (problèmes rencontrés dans une situation de vie réelle).

Etant donné le nombre de tâches nécessitant la vision dans le monde actuel, les restrictions et limitations occasionnés par une déficience visuelle sont évidemment nombreuses. Depuis le début des années 1990, un nombre important de travaux a tenté d'évaluer les conséquences de la déficience visuelle sur la vie de tous les jours, en utilisant des indicateurs de dépendance, ou plus récemment des outils de mesure de la qualité de vie. Les incapacités relevées dans ces différentes enquêtes dépendent naturellement fortement de l'origine et de la nature du handicap, qu'il s'agisse d'un déficit d'acuité, de champ visuel, ou de sensibilité au contraste, mais également du profil des personnes atteintes, notamment selon l'âge auquel ces symptômes sont apparus.

Ainsi, chez les personnes âgées, de nombreuses études ont montré l'incidence des troubles visuels sur les chutes [Klein et al., 1998; Rq et al., 1998], les fractures de la hanche [Felson et al., 1989; Ivers et al., 2000], le placement en maison de retraite [Mitchell et al., 1997; Mormiche and Boissonnat, 2003; Wang et al., 2003], le recours aux services d'aide à la personne [Wang et al., 1999] ou encore le taux de mortalité [Christ et al., 2008; Jacobs et al., 2005; Wang et al., 2001].

D'une manière plus générale, les activités les plus affectées par les déficiences visuelles sont la communication écrite et la cognition spatiale (compréhension de

l'environnement, localisation d'objets, navigation). Ces incapacités entraînent de nombreux handicaps, ou désavantages, selon la terminologie de l'OMS, dans plusieurs grands domaines sociétaux tels que l'accessibilité à l'information, l'inclusion dans les sphères professionnelle ou associative, l'accès à la culture et aux loisirs, etc.

La déficience visuelle induit donc une baisse globale de la qualité de vie [Valbuena et al., 1999; Wahl et al., 1999; West et al., 2002]. L'étude américaine de la cohorte EPESE[1] a par exemple montré qu'une acuité visuelle inférieure à 1/10 divise par plus de quatre les activités quotidiennes [Salive et al., 1994]. Les conséquences de la malvoyance et de la cécité touchent à la fois à l'autonomie [Rubin et al., 2001; Varma et al., 2006; Whitson et al., 2007], à la mobilité [Friedman et al., 2007; Geruschat and Turano, 2007; Turano et al., 2004, 1999], à l'état de santé général [Jacobs et al., 2005; Wallhagen et al., 2001; Wang et al., 2000], et sont aussi associées à des troubles psychologiques [Ip et al., 2000; Lee et al., 2000]. En effet, en plus des aspects physiques évidents, les répercussions des déficiences visuelles peuvent être émotionnelles et sociales, causant par exemple isolement et dépressions [Carabellese et al., 1993; Chia et al., 2004; Wallhagen et al., 2001].

[1] Established Populations for the Epidemiologic Studies of the Elderly

2. Systèmes d'assistance visuelle basés sur une approche holistique

Lorsque les causes de malvoyance ou de cécité ne sont pas traitables et que le déficit restera permanent, comme dans le cas de la dégénérescence maculaire, des glaucomes ou de nombreuses opacités de la cornée, différentes solutions ont été proposées pour compenser la perte visuelle, et restaurer, si ce n'est un réel sens visuel, au moins certaines des fonctions perdues. Deux catégories de systèmes se démarquent. D'une part, les approches holistiques, visant à restituer l'information visuelle dans sa globalité. C'est la démarche commune des systèmes de substitution sensorielle et des neuroprothèses que nous développerons ici. D'autre pat des aides spécifiques, tentant de répondre aux besoins identifiés dans des tâches spécifiques, qui seront abordés dans la section suivante.

2.1 Substitution sensorielle

Les systèmes de substitution sensorielle sont des dispositifs qui permettent que des informations normalement acquises par un organe sensoriel défaillant soient restituées vers une autre modalité perceptive. Dans le cas des aveugles ou des malvoyants, ils consistent par exemple à transmettre des informations visuelles via le système auditif ou somesthésique. Ces systèmes reposent tous sur une architecture en 3 étapes, commençant par l'acquisition de l'information (traditionnellement une caméra embarquée), puis la conversion et/ou le traitement de celle-ci pour la retransmettre à la modalité sensorielle de sortie, et finalement par sa restitution via un dispositif adapté.

2.1.1 Substitution visuo-tactile

Le Braille, inventé en 1929 et toujours largement employé de nos jours, constitue une des premières tentatives visant à présenter des informations de nature visuelle (en l'occurrence du texte) sous forme tactile. Le système Braille, composé de cellules de 6 points, permet de coder 64 combinaisons sur chacune d'elles, et donc de représenter l'ensemble des lettres de l'alphabet, ainsi que les chiffres, les caractères spéciaux propres à chaque langue, et même les notations musicales ou mathématiques.

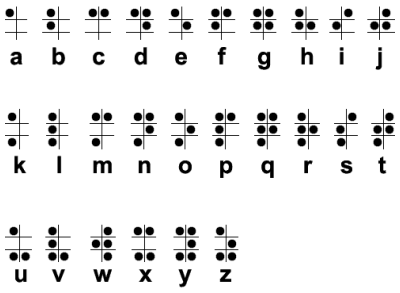


Figure I-8 Alphabet Braille.

Le Braille en tant que tel ne peut être considéré comme de la substitution sensorielle, car il n'est pas possible de convertir directement de l'information visuelle en codage Braille, néanmoins différents outils complémentaires permettent de lui apporter cet aspect dynamique. Il existe par exemple de nombreux logiciels appelés lecteurs d'écrans, permettant de retranscrire à la volée le contenu d'un écran sous forme parlée, ou par le biais d'une plage Braille telle que celle présentée dans la Figure I-9. Il est également possible d'utiliser des systèmes de reconnaissance de caractères permettant de convertir en texte un document scanné ou une image, qui à son tour pourra être restitué en langage Braille. Un des premiers dispositifs reposant sur ce principe a été commercialisé en 1971[1] par la compagnie américaine *Telesensory System*. Baptisé Optacon, il consistait en un stylet équipé d'une caméra que l'utilisateur déplaçait le long du texte, et d'une matrice de picots vibrants reproduisant la forme des caractères [Goldish and Taylor, 1974]. Ce système ne reposait pas sur le codage braille, ni sur la reconnaissance automatique de caractères, mais directement sur la forme des lettres et des motifs transposée sur une matrices de 6 colonnes et 24 rangées (visible sur la Figure I-9). L'utilisation de ce dispositif nécessitait un long entraînement, et la vitesse de lecture restait relativement faible y compris pour un utilisateur expérimenté, entre 20 et 80 mots à la minute.

Figure I-9 A gauche, plage braille permettant la lecture sur ordinateur ; à droite, dispositif Optacon.

Les approches de réelle substitution visuo-tactile, ne se limitant pas à la lecture, ont été initiées par Paul Bach-y-Rita et ses collaborateurs dans les années 70. Le dispositif proposé, appelé Tactile Vision Substitution System (TVSS), convertissait des informations visuelles capturées par une caméra en des sensations tactiles appliquées à la surface du corps [Bach-y-Rita et al., 1969a, 1969b]. Dans le premier dispositif, une grille de stimulation électro-tactile était montée sur une chaise de dentiste afin de stimuler le dos du sujet, et la caméra utilisée était fixe, comme montré dans la Figure I-10.

[1] Puis sa production arrêtée en 1971 pour raisons économiques.

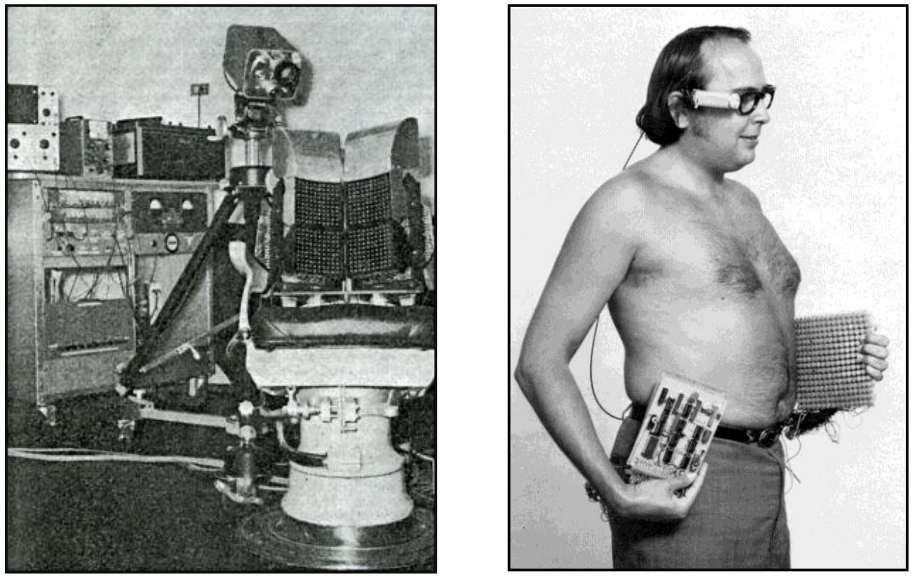


Figure I-10 Tactile Vision Substitution System (TVSS) de Paul Bach-Y-Rita.

Par la suite, différents prototypes se sont succédés. Le deuxième, également présenté dans la Figure I-10, ne stimulait non plus le dos mais l'abdomen [Bach-y-Rita, 1983]. La différence majeure résidait dans le fait que la caméra était mobile et manipulée par l'utilisateur. Différentes expériences ont effectivement montré que ce contrôle était nécessaire à l'utilisation d'un système de substitution sensorielle [Arno et al., 2001a; Auvray and Myin, 2009; Bach-y-Rita, 2002; Guarniero, 1974]. Non seulement les performances de discrimination de formes s'en trouvent très largement augmentées (par rapport à une caméra fixe ou actionnée par une autre personne que le sujet), mais la nature même de la perception rapportée par l'utilisateur s'avère radicalement différente. Les sujets témoignent de ce changement de perception [Bach-y-Rita, 1983] :

> *When the camera was either immobile or under the control of another person the subjects reported experiences in terms of sensations on the area of skin which was receiving the stimuli. However, when they could easily direct the camera at will, their reports were in terms of objects localized externally in space in front of them. The provision of a motor linkage (camera movement) for the sensory receptor surface on the skin produced a surrogate "perceptual organ".*

La manipulation de la camera permet non seulement la mise en place d'une boucle sensori-motrice nécessaire à l'extériorisation de la perception, mais également l'enrichissement des informations acquises dans le cadre des premiers dispositifs à résolution limitée (du fait du nombre d'éléments des matrices de stimulation et de la faible qualité des jugements tactiles dans les régions du dos ou de l'abdomen). La quantité d'information perçue à un instant donné étant trop faible pour l'interprétation de la scène,

les utilisateurs se trouvaient contraints de balayer l'espace pendant 30 à 60 secondes afin d'identifier les objets présents, et ce en se basant sur les changements de contours résultant du mouvement de la caméra.

Un dernier type de dispositif fut développé à la fin des années 90 par l'équipe de Bach-Y-Rita, pour compenser cette faible acuité somesthésique au niveau du dos ou de l'abdomen. Baptisé Tongue Dispay Unit (TDU), il consistait en une matrice de stimulation de 49 électrodes disposée sur la langue. Un système très similaire, mais appliqué sur le palais, a été développé plus tard par Tang et Beebe [Tang and Beebe, 2006, 2003]. En effet, la cavité buccale est un des organes ayant la plus forte densité de récepteurs tactiles, ce qui permet une plus grande résolution de perception et demande moins d'énergie de stimulation[1]. La surface corticale dédiée à la langue est à titre d'exemple plus grande que celle dédiée à toute la surface du dos, comme illustré dans la Figure I-11.

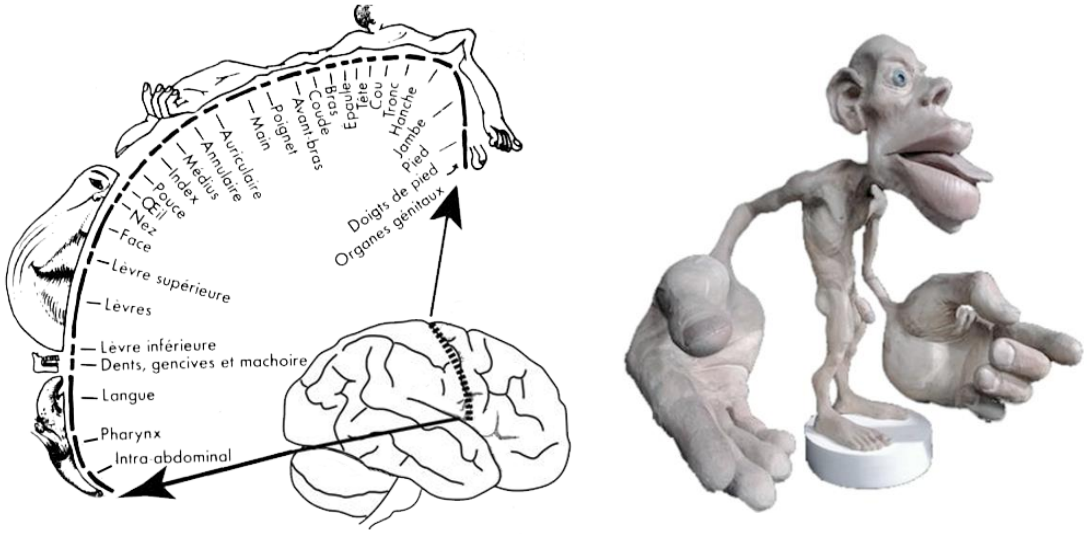


Figure I-11 Homonculus somesthésique, indiquant la part relative des zones corticales allouées à chaque partie du corps.

La première version du TDU avait une résolution de 7x7 électrodes de stimulation et permettait la reconnaissance de formes simples, en l'occurrence des ronds, carrés et triangles [Bach-y-Rita et al., 1998]. Ces résultats sont comparables aux résultats rapportés dans une autre étude [Kaczmarek et al., 1997] avec le même protocole et le même dispositif expérimental, mais appliqué sur le bout d'un doigt. Les performances de reconnaissance des motifs sont alors très proches (90% environ pour des grandes tailles de stimuli) de celles obtenues avec une stimulation de la langue.

[1] Seulement 5 à 15 V et 0.4 à 2.0 mA (Paul Bach-Y-Rita & Stephen W. Kercel, 2003).

Au début des années 2000, une nouvelle version du TDU (voir Figure I-12) est créée avec une matrice de 144 électrodes de stimulation (12x12) connectée à une caméra de faible résolution (240x180) et de 54° d'angle de vue [Sampaio et al., 2001]. L'acuité "visuelle" avec un tel système a été mesurée grâce au test standard de Snellen. Deux groupes de sujets n'ayant jamais utilisé ce type de dispositif ont été constitué, l'un comportant 6 voyants et l'autre 6 non-voyants congénitaux. Les stimuli étaient dérivés du 'E' de Snellen, dans six tailles (5 ; 3,6 ; 2,5 ; 1,8 ; 1,5 et 0,85 cm) et quatre orientations différentes. Les sujets pouvaient faire bouger manuellement la caméra, fixée à 40 cm de la source par un bras articulé. Avant tout apprentissage du système, l'acuité des sujets était proche de 20/860 (seuls les plus grands stimuli, de 5*5 cm, étaient différenciables), et similaire dans les deux groupes. Après un apprentissage de neuf heures consistant en la détection de lignes de taille et d'orientation différentes, l'acuité visuelle avait doublé (20/430) mais restait néanmoins très faible.

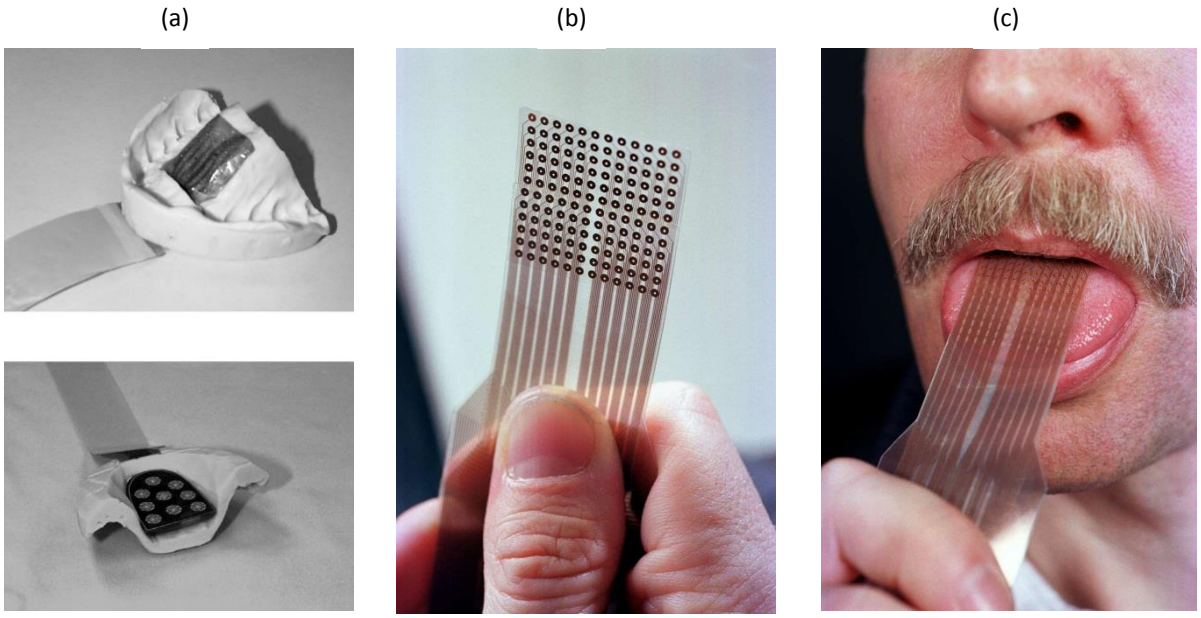


Figure I-12 (a) Dispositif de (Tang and Beebe, 2006, 2003) composé d'une matrice de stimulation du palais et d'un pavé tactile actionnable par la langue ; (b) et (c) Tongue Display Unit (TDU) de Bach-Y-Rita.

2.1.2 Substitution visuo-auditive

Apparus plus tard, d'autres systèmes pour non-voyants utilisent l'audition plutôt que le toucher pour restituer l'information visuelle. Le système auditif montre en effet des seuils de discrimination d'intensité, de fréquence et de position extrêmement fins, et il est capable de traiter des motifs sonores complexes et changeants comme la parole même dans des

environnements bruyants [Hirsh, 1988]. La plupart des aveugles utilisent d'ailleurs les sons pour la navigation, ceux du bruit de leurs pas ou de leur canne sur le sol, ainsi que ceux de leur voix ou des activités de la ville, qui les informent sur l'environnement et les obstacles susceptibles d'être rencontrés. Ces capacités auditives d'écholocalisation ont été observées dès 1947 dans [Worchel and Dallenbach, 1947]. De plus, systèmes de substitution visuo-auditive ont l'avantage de consommer peu d'énergie et d'utiliser des technologies courantes, compactes et peu coûteuses (généralement une caméra/webcam, des écouteurs et un ordinateur portable ou un téléphone). Néanmoins, les images numériques étant par nature bidimensionnelles, la transcription des informations visuelles en informations auditives est donc plus délicate que dans le cas de la substitution visuo-tactile. Une interface tactile possède en effet une structure spatiale bidimensionnelle sur laquelle on peut simplement recopier l'image, ce qui n'est pas le cas d'une interface sonore. La substitution de la vision par l'audition s'obtient en transformant l'image vidéo provenant d'une caméra en un signal sonore complexe transmis via des écouteurs, en utilisant quatre propriétés des sons: la fréquence, l'intensité, le délai et les différences inter-aurales.

Un grand nombre de prototypes reposant sur la substitution sensorielle ont été proposé depuis les années 90 (et particulièrement après les années 2000, voir par exemple la revue de [Maidenbaum et al., 2014]), mais nous ne présenterons ici que quatre d'entre eux, ayant connu un fort succès[1] :

1. Le plus connu, 'The vOICe' (les lettres capitales étant lues « Oh I see ») [Meijer, 1992], est développé depuis 1992 par l'ingénieur Peter Meijer, au sein du laboratoire Philips Research à Eindhoven.

2. Le système PSVA [Arno et al., 1999] a été mis au point en 1999 par Capelle et ses collaborateurs à l'Université Catholique de Louvain.

3. Un système plus récent, 'The Vibe' [Durette et al., 2008], est issu d'une collaboration entre le laboratoire de Neurophysique et Physiologie du Système Moteur (Sylvain Hanneton) et le laboratoire de Psychologie Expérimentale, tous deux à l'Université René Descartes de Paris (Sylvain Haupert, J. Kevin O'Regan, Malika Auvray).

4. Enfin, le système See ColOr (Seeing Colors with an Orchestra), conçu à l'université de Genève par Thierry Pun, Guido Bologna et leur groupe [G. Bologna et al., 2009; Bologna et al., 2007], qui fait toujours l'objet de recherches actives et a profité de nombreuses évolution au cours des dernières années [Gomez Valencia, 2014].

[1] Du moins dans la communauté scientifique, car comme nous le verrons, l'usage de ces systèmes ne s'est pas répandu dans la population non-voyante.

Les trois premiers présentent des architecture relativement semblables, et diffèrent principalement par leur codage de l'information : le premier, the vOice, repose sur un balayage séquentiel de l'image, alors que les deux suivants sont « simultanés », chaque image étant retransmise sous la forme d'un seul son complexe. Le dernier, d'un concept assez différent, sera détaillé un peu plus loin.

The vOice

Dans le système The vOICe, l'image, convertie en niveau de gris et à une résolution de 64x64 pixels, est balayée horizontalement toutes les secondes. Chaque colonne est donc successivement représentée par un son d'une durée de 1/64ème de seconde, soit environ 15 ms. La position verticale est quant à elle codée par une fréquence. Plus le pixel est haut, plus le son est aigu. L'intensité du pixel, finalement, est représentée par l'intensité du son produit sur sa fréquence correspondante, comme illustré dans la Figure I-13, tirée de [Meijer, 1992].

Différents prototypes utilisent le système de codage de The vOICe, comme celui présenté dans la Figure I-13, où la caméra est montée sur un casque. Ce type de dispositif au design futuriste n'est cependant pas très bien reçu par la communauté des non-voyants pour un usage quotidien, ceux-ci préférant généralement un équipement plus neutre, n'attirant pas l'attention. La majeure partie des utilisateurs se tourne donc plutôt vers des lunettes équipées d'une petite caméra centrale et d'écouteurs classiques connectés à un mini-PC. Récemment une version Android de l'application a même été développée, permettant l'utilisation de The vOICe sur la plupart des téléphones actuels, ainsi que sur les toutes nouvelles Google Glass.

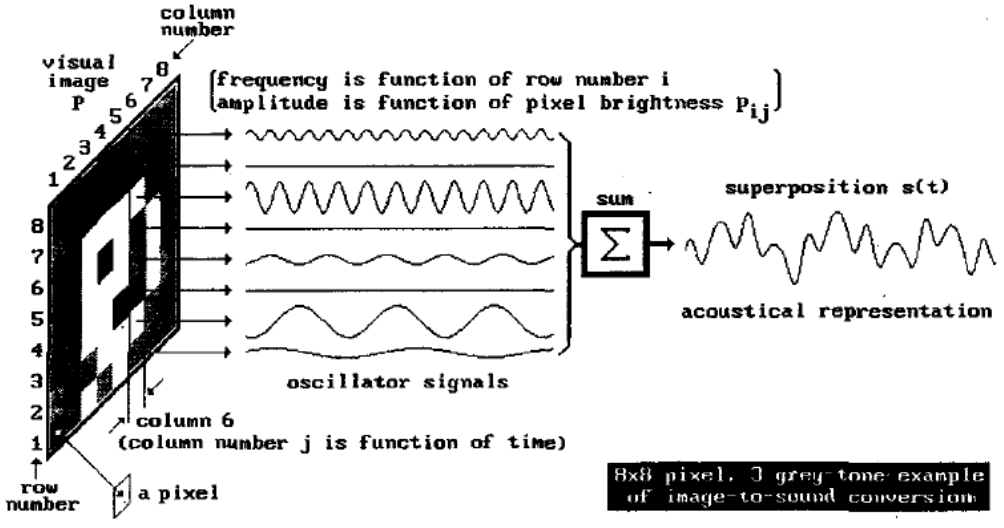


Figure I-13 Principes de la conversion d'une image en son réalisée par The vOICe pour un exemple d'image de 8 par 8 pixels avec 3 niveaux de gris.

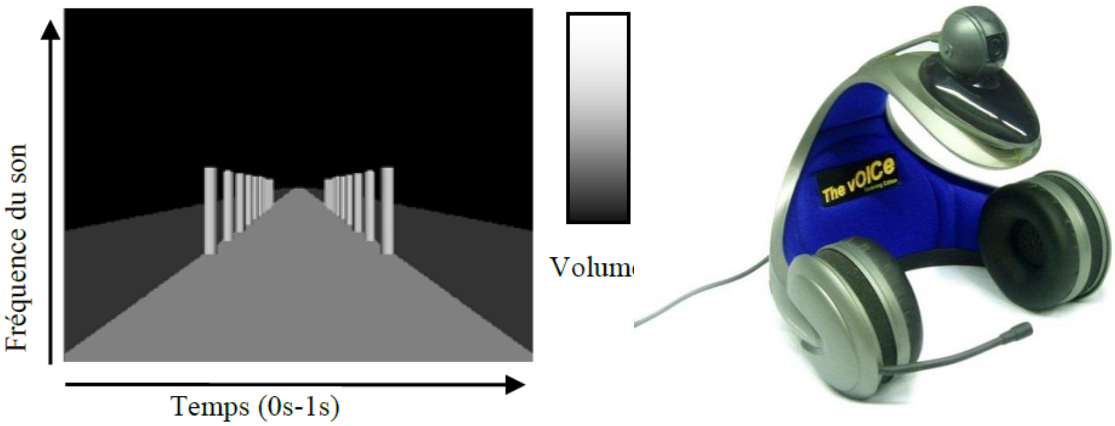


Figure I-14 The vOICe. A gauche illustration de l'encodage d'une image en fonction de la position et de l'intensité de chaque pixel ; A droite prototype du dispositif intégrant webcam, écouteurs et micro.

The vOICe a été évalué dans des tâches de reconnaissance d'objets au cours de la thèse de Malika Auvray [Auvray, 2004; Auvray et al., 2005]. Les objets étaient disposés sur une table blanche située à environ un mètre des sujets. Les résultats présentés montrent que les sujets, équipés d'une caméra à la main ou sur la tête, pouvaient localiser et reconnaître les objets avec des temps moyens avoisinant 50 secondes (voir Figure I-15). Des travaux complémentaires [Auvray et al., 2007], toujours dans des environnements contrôlés, ont aussi montré qu'il était possible d'entraîner les sujet équipés du dispositif à accomplir des tâches de discrimination et des mouvements de pointage avec une erreur moyenne de 6 à 15 cm en fonction de la distance de l'objet.

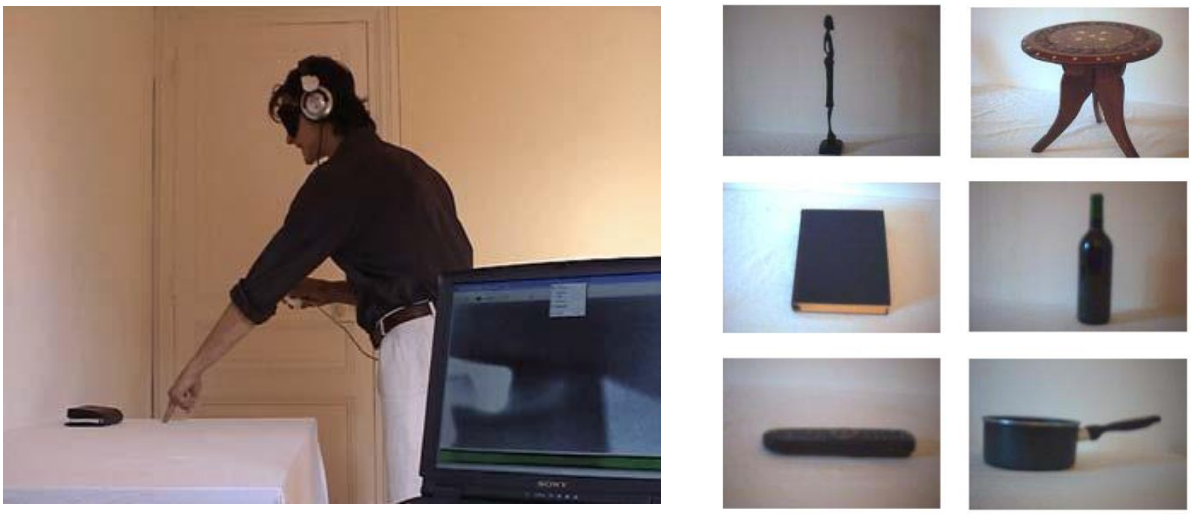


Figure I-15 Evaluation du système The vOICe'. A gauche : sujet réalisant une tâche de déplacement et de pointage. A droite : exemples d'objets utilisés pour la tâche de reconnaissance.

PSAV

Développé à partir du modèle théorique présenté dans [Veraart, 1989], Claude Veraart et son groupe de l'Université catholique de Louvain, proposèrent en 1998 un dispositif similaire à 'the vOICe', nommé PSVA (Prosthesis for Substitution of Vision by Audition) [Capelle et al., 1998]. Il repose sur l'idée que, une fois passée la phase d'encodage des stimuli en signal neural, il existe une similarité entre le traitement de l'information par le système visuel et par la modalité de substitution. Il propose donc d'appliquer aux images captées par les caméras, un filtrage par un modèle simplifié de rétine, puis par un modèle du cortex visuel primaire, et enfin par un modèle inverse du système auditif et de la cochlée. Plus concrètement, ce traitement consiste en une détection de contours basée sur le Laplacien de Gaussienne [Marr, 1982] et en une organisation multi-échelle reproduisant le principe de vision fovéale et périphérique. Celle-ci est représentée dans la Figure I-16, où l'on observe que dans la zone centrale les points sont 16 fois plus denses que dans le reste de l'image. La conversion de l'image en son repose, elle, sur différentes propriétés du système auditif :

- Les fréquences doivent être comprises entre 50hz et 15kHz pour être audibles par la majorité des utilisateurs.
- Les fréquences voisines doivent respecter les seuils de discrimination moyens.
- Les fréquences choisies doivent être distribuées selon une échelle exponentielle pour être perçues comme équidistantes.
- La discrimination horizontale résulte des différences inter-aurales de phase et d'intensité, alors que verticalement elle repose sur la modification du spectre causée par les réflexions dans le pavillon de l'oreille.

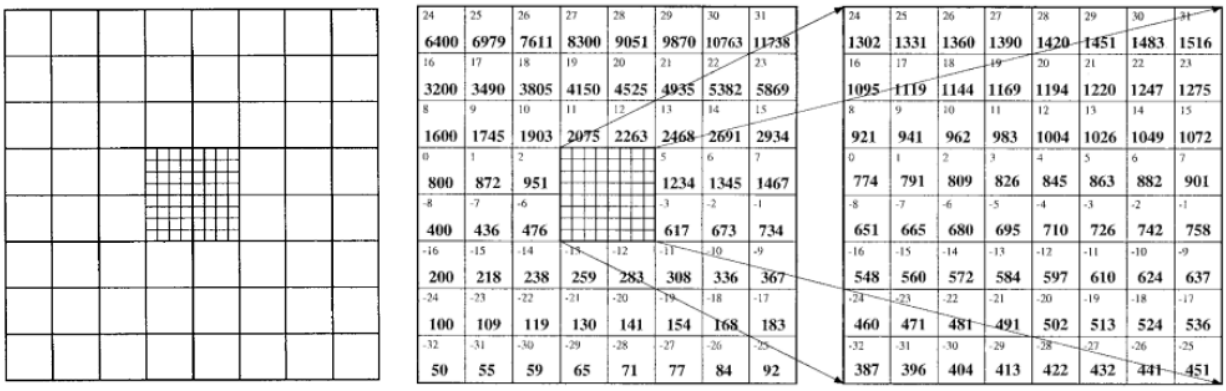


Figure I-16 Encodage de l'image du système PSAV. A chaque résolution est indiqué en haut le numéro du pixel (de -32 à +31), et en bas la fréquence correspondante. Celle-ci est calculée en périphérie par la formule $f = f_c \times 2^{(2p+1)/16}$ et au centre par $f = f_c \times 2^{(2p+1)/64}$ où la fréquence centrale est $f_c = 766.08$ hz [Capelle et al., 1998].

Par conséquent, l'encodage retenu retransmet la position verticale en hauteur de son (aigus en haut de l'image, graves en bas), la position horizontale par la différence binaurale, et le niveau de gris par l'intensité du son. A chacun des pixels est donc associée une fréquence spécifique, comme illustré dans la Figure I-16. Des jeux de sons différents sont choisis pour la partie centrale et périphérique de l'image, et les fréquences de pixels voisins sont choisies de sorte que les colonnes produisent des sons mélodieux, et les lignes des harmonies dissonantes.

Ce système offre l'avantage d'être « immédiat », contrairement à 'The vOICe' qui nécessite une seconde pour le balayage horizontal de l'image. Le PSAV peut donc atteindre des vitesses de rafraîchissement de 25hz, bien que, pour des raisons pratiques, il a finalement été cadencé à 10hz. Ce gain permet une vitesse de rafraîchissement suffisante pour la perception du mouvement. Néanmoins, le choix d'un encodage fréquentiel des positions verticales et horizontales limite sa résolution à 124 positions, pour plus de 1000 dans le cas de 'The vOICe'.

Vingt-quatre sujets voyants, âgés de 19 à 36 ans et ne présentant pas de déficit d'audition ont été répartis en deux groupes de 12 personnes (le groupe contrôle et le groupe test), puis ont évalué le système PSAV les yeux bandés [Capelle et al., 1998]. Le groupe test a effectué 14 sessions : 10 sessions d'apprentissage et 4 pour l'évaluation du système (une au début, deux au milieu puis une à la fin des sessions d'apprentissage), tandis que les sujets contrôles ont seulement effectué les 4 sessions d'évaluation. Durant les sessions d'apprentissage, les sujets du groupe test devaient apprendre à reconnaître les stimuli grâce au système PSAV et au feedback des expérimentateurs. Durant les sessions d'évaluation, aucun feedback n'était donné. 50 stimuli dérivés de 15 motifs visuels simples présentés dans différentes orientations (voir la Figure I-17) ont été évalués dans des tâches de localisation et de reconnaissance. Les résultats montrent que s'il y a un effet significatif de l'apprentissage sur les performances et sur le temps de reconnaissance, ceux-ci restent cependant élevés dans les deux groupes de sujets (entre 1 et 2 minutes par motif) [Arno et al., 1999].

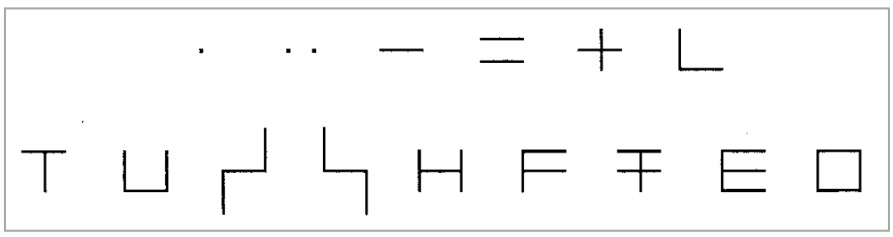


Figure I-17 Stimuli utilisés dans la tâche de reconnaissance de motifs visuels avec le dispositif PSVA.

The Vibe

The Vibe a été développé en 2004 au sein de l'Université René Descartes par le laboratoire de Neurophysique et Physiologie du Système Moteur (Sylvain Hanneton) et le laboratoire de Psychologie Expérimentale (Sylvain Haupert, J. Kevin O'Regan, Malika Auvray) [Auvray, 2004; Auvray et al., 2005; Hanneton et al., 2010]. Similaire aux systèmes précédents, il propose d'encoder la position verticale du pixel en fréquence (qui augmente avec la position verticale du pixel) et les niveaux de gris en intensité du son. La position horizontale est ici encodée en disparité binaurale, c'est-à-dire par les différences inter-aurales de temps et d'intensité.

Ce système repose par ailleurs sur la segmentation de l'image en champs récepteurs (voir Figure I-18, tirée de [Durette et al., 2008]), dont le nombre, les positions et le recouvrement sont configurables. Un champ récepteur est un groupe de pixels auquel est attribué un son élémentaire en fonction de la position de son barycentre et du niveau de gris moyen (ou tout autre fonction des intensités de ces pixels). La répartition des champs récepteurs est, dans cet exemple, obtenue par un algorithme d'auto-organisation de Kohonen, mais elle peut également suivre une grille régulière. Le son final est produit en sommant les sons des différents champs récepteurs de l'image. Comme pour le système PSVA, ce son complexe retranscrit l'ensemble de l'image, un balayage n'est donc pas

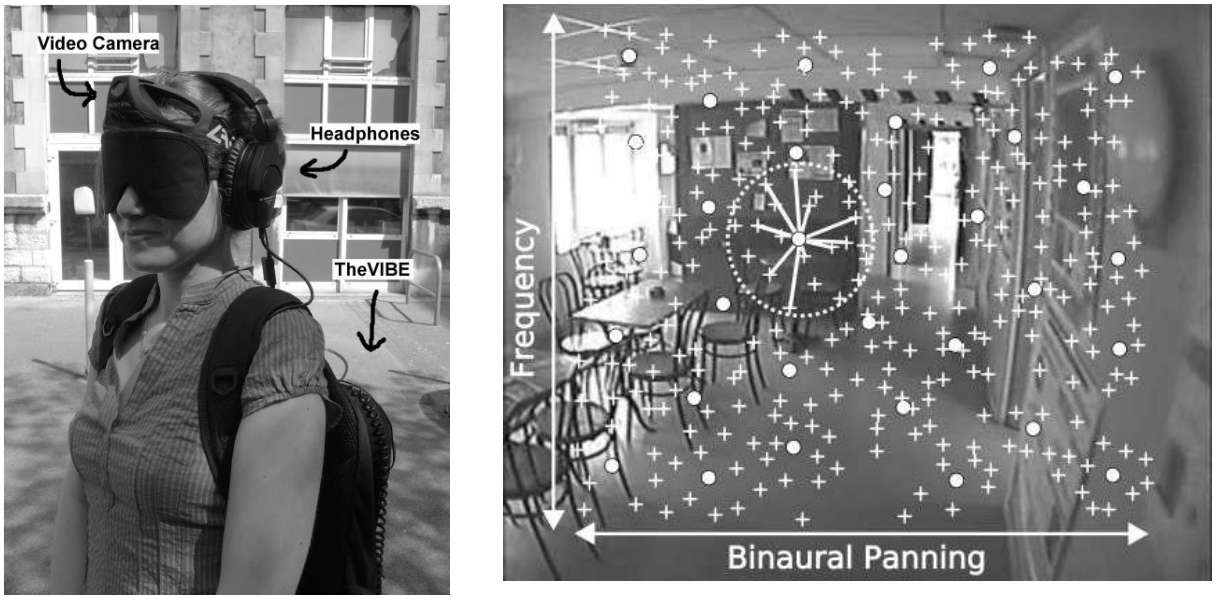


Figure I-18 The Vibe. A gauche, prototype du dispositif comprenant une caméra vidéo montée sur le le front, des écouteurs et un ordinateur portable dans le sac ; A droite, représentation de l'encodage de l'image en sons. Des ensembles de pixels (représentés par les croix) sont regroupés au sein de champs récepteurs (les points). A ceux-ci sont associés une intensité (moyenne des pixels associés) ainsi qu'une fréquence et une balance binaurale en fonction de sa position.

nécessaire. Ils sont donc présentés de façon continue, la durée de présentation de ce son complexe étant égale à la période de rafraîchissement de la vidéo, fixée usuellement à 25 images par secondes.

Le système 'The Vibe' a été évalué dans différentes tâches [Durette, 2009]. Durette et collaborateurs ont tout d'abord demandé à seize sujets voyants ayant les yeux bandés, et équipés du dispositif, de localiser puis d'aller saisir une cible lumineuse fixée à différents endroits dans une salle sombre, tel qu'illustré dans la Figure I-19. Le temps moyen pour aller chercher cette cible diminuait au cours de l'apprentissage jusqu'à atteindre 20 secondes. Une autre tâche de pointage a été réalisée pour valider la précision de la localisation dans [Hanneton et al., 2010].

Figure I-19 Tâche de localisation et de préhension d'une cible à l'aide du dispositif The Vibe.

En vue d'évaluer l'utilité de ce dispositif pour l'aide à la mobilité, une étude a été réalisée dans une tâche de navigation en environnement réel [Durette et al., 2008]. Vingt sujets voyants ayant les yeux bandés portaient sur la tête une caméra grand angle (92°) et devaient effectuer un parcours en 'U' durant quatre sessions expérimentales espacées d'au moins 24h. Le temps total pour effectuer le trajet et le nombre de collisions étaient enregistrés. Durant l'apprentissage, réalisé au cours des trois premières sessions, l'expérimentateur guidait le sujet avec le bras (première session) ou verbalement par des instructions « gauche » ou « droite » (pour les sessions 2 et 3, voir Figure I-20). En moyenne, le temps de parcours diminuait, passant de 235 secondes lors de la première session à 170 secondes lors de la troisième. De même, le nombre de collisions durant le trajet passait en moyenne de 7,2 à 6 au cours des trois sessions. La condition contrôle, qui consistait à inverser l'image verticalement, a montré que le dispositif orienté normalement permettait de réaliser la tâche plus rapidement.

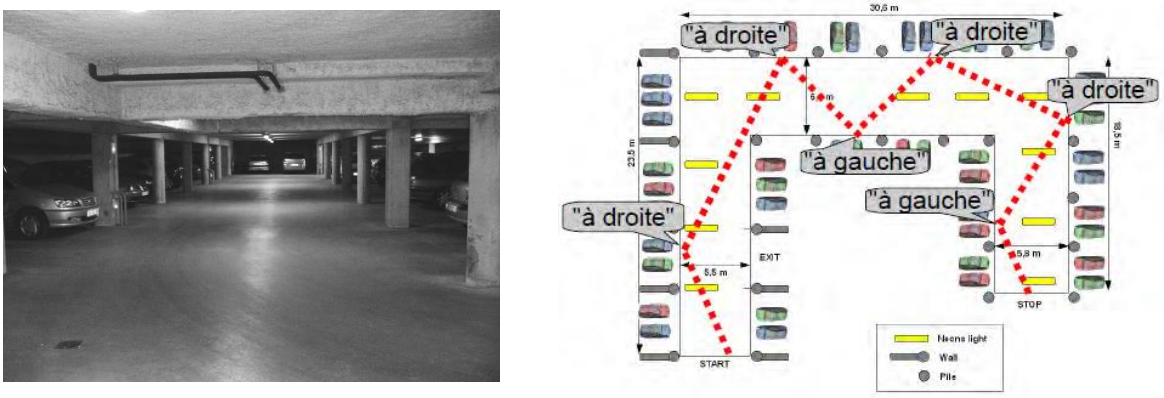


Figure I-20 Evaluation de The Voice dans une tâche de mobilité. A gauche, photo du parking du laboratoire où est réalisé le trajet, offrant un terrain large, sans obstacles et d'éclairage constant ; A droite, exemple d'un trajet réalisé dans la phase d'apprentissage avec les instructions de l'expérimentateur associées.

See ColOr

Alors que la majorité des systèmes de suppléance sensorielle se basent seulement sur l'intensité des pixels dans des images en niveaux de gris, See ColOr, tel que son nom l'indique, fournit à l'utilisateur des informations visuelles basées sur la couleur. Il repose sur la synthèse de sons binauraux spatialisés et sur des caméras stéréoscopiques[1]. Chaque pixel est représenté par une source sonore directionnelle. La teinte chromatique se traduit par un instrument, la saturation par une hauteur des notes, et la distance par la durée du son (voir Figure I-21).

Dans les premières versions du système, seule une ligne horizontale était sonifiée, discrétisée en 25 points[2] répartis dans toute la largeur du champ visuel des caméras. Ce choix s'explique par une mauvaise perception de l'angle vertical des sons binauraux spatialisés au moyen d'HRTF génériques. Par la suite de nombreuses variantes du système ont été développées, incluant par exemple une tablette [Bologna et al., 2011]. Celle-ci permet de remplacer les balayages de la tête nécessaire à l'exploration de l'espace par des mouvements des doigts sur la tablette.

[1] Les dernières versions du prototype ont remplacé la stéréovision par le couplage d'une webcam et d'une caméra Time-of-Flight, permettant de calculer des cartes de profondeurs alignées avec une image RGB.
[2] Ceux au centre sont plus ressérés qu'en périphérie, d'une manière similaire à l'organisation du système visuel.

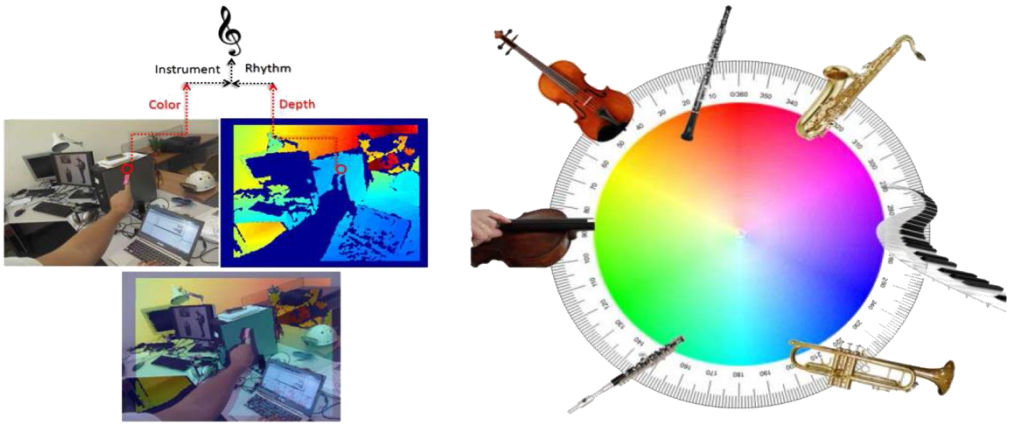


Figure I-21 Encodage des sons dans le système See ColOr: la distance d'un pixel (calculée par stéréovision) est codée par le rythme, la teinte chromatique par un instrument de musique, et la saturation par la hauteur (ou *pitch*) de la note.

Par la suite, de nombreuses fonctionnalités ont été ajoutées au dispositif, comme la sonification d'un compas, ou l'exploration de cartes de profondeurs (les couleurs et palettes d'instruments ne sont pas utilisées dans ce mode, à la place, une interface sonore assez simple permet de signaler à l'utilisateur la présence ou l'absence d'obstacles dans la zone explorée correspondant au point de la tablette touché). Un signal d'avertissement de collision a aussi été ajouté pour indiquer des dangers potentiels quand le nombre de points distants de moins d'un mètre augmente dans la carte de profondeur [Bologna et al., 2011], ou encore un module de détection de texte et de reconnaissance automatique de caractères [Gomez Valencia, 2014].

Dans son utilisation classique (le mode nommé « Local Perception » [Bologna et al., 2011]), le système See ColOr a pu permettre la réalisation de différentes tâches visuelles de navigation, de localisation ou de reconnaissance (telles que se déplacer le long d'une ligne au sol tracées au sol [Guido Bologna et al., 2009], différencier certains fruits, localiser des sources lumineuses, ou encore assortir ses paires de chaussettes par couleur [Bologna et al., 2008]), bien que réalisées dans des temps relativement lents, à l'instar des autres systèmes de substitution sensorielle (une dizaine de minutes en moyenne pour assortir 5 paires de chaussettes, contre 25 secondes pour un voyant.

En revanche les nombreuses évolutions récentes le démarquent clairement des approches holistiques classiques [Deville et al., 2009; Gomez Valencia, 2014]. En effet, plutôt que de reposer sur un système figé de conversions de la scène visuelle en une « scène auditive », See ColOr s'est enrichi de nombreuses méthodes avancées de vision artificielle (comme l'utilisation de cartes de saillances, de reconnaissance automatique de caractères,

de suivi d'objet, d'analyse d'obstacle, de segmentation de l'image, de détecteur de collision, etc...) permettant d'ouvrir un large éventail de nouvelles possibilités prometteuses. Dans ce sens il se trouve à cheval avec la deuxième catégorie de systèmes que nous aborderons dans la section I.3, les dispositifs d'aide basés sur une approche fonctionnelle.

Figure I-22 Utilisation du dispositif See ColOr dans différentes tâches visuelles (suivi d'une ligne de couleur, appariement de chaussettes de même couleur, recherche de la sortie d'une salle, ou encore d'un teeshirt rouge)

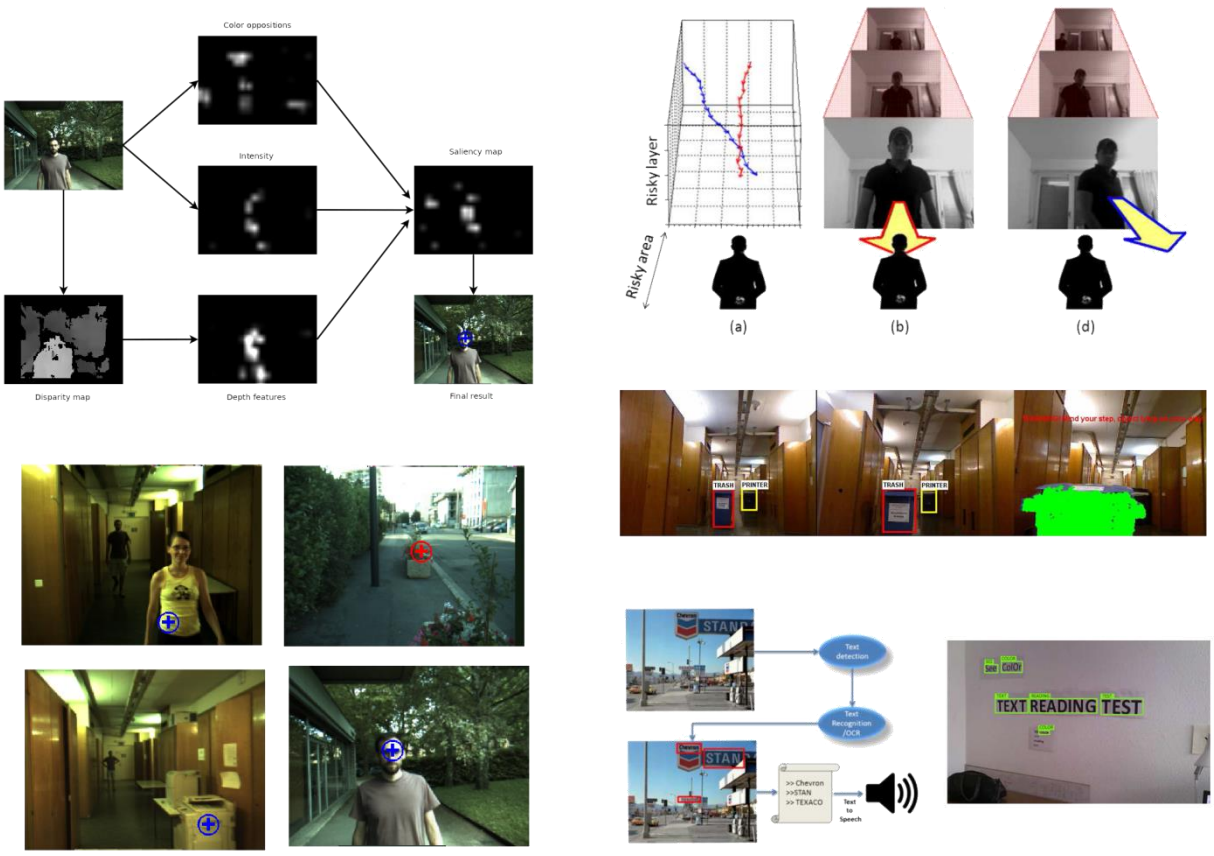


Figure I-23 Apercu de la vision artificielle dans le projet See ColOr. A gauche carte de saillances permettant d'identifier des zones d'intérêt de l'image. A droite, reconnaissance d'objets, d'obstacles et de dangers ; détection et reconnaissance automatique de caractères. (se reporter à [Gomez Valencia, 2014], dont sont extraites les figures, pour des détails)

2.1.3 Plasticité cérébrale

Si les différents systèmes de substitution sensorielle existants ont pour but premier d'aider les aveugles dans leur vie quotidienne, leur basse résolution, leur faible efficience[1], et différents inconvénients sur lesquels nous reviendrons par la suite rendent leur usage souvent inadapté. Ces raisons expliquent qu'extrêmement peu de non-voyants n'utilisent à ce jour ce type de systèmes.

Cependant, dans le cadre de la recherche fondamentale, les dispositifs de substitution sensorielle se sont avérés être de puissants outils pour l'étude de la plasticité du cerveau ou du développement chez l'enfant déficient visuel [Aitken and Bower, 1983], et ont donc fait l'objet de nombreux travaux dans le domaine des neurosciences. En raison de la privation sensorielle, certaines aires corticales visuelles sont « recrutées » par d'autres modalités chez les aveugles (voir Figure I-24)Figure I-25. En introduisant l'utilisation de systèmes de substitution il devient alors possible d'observer les réorganisations fonctionnelles induites au cours de l'apprentissage [Dagnelie, 2011].

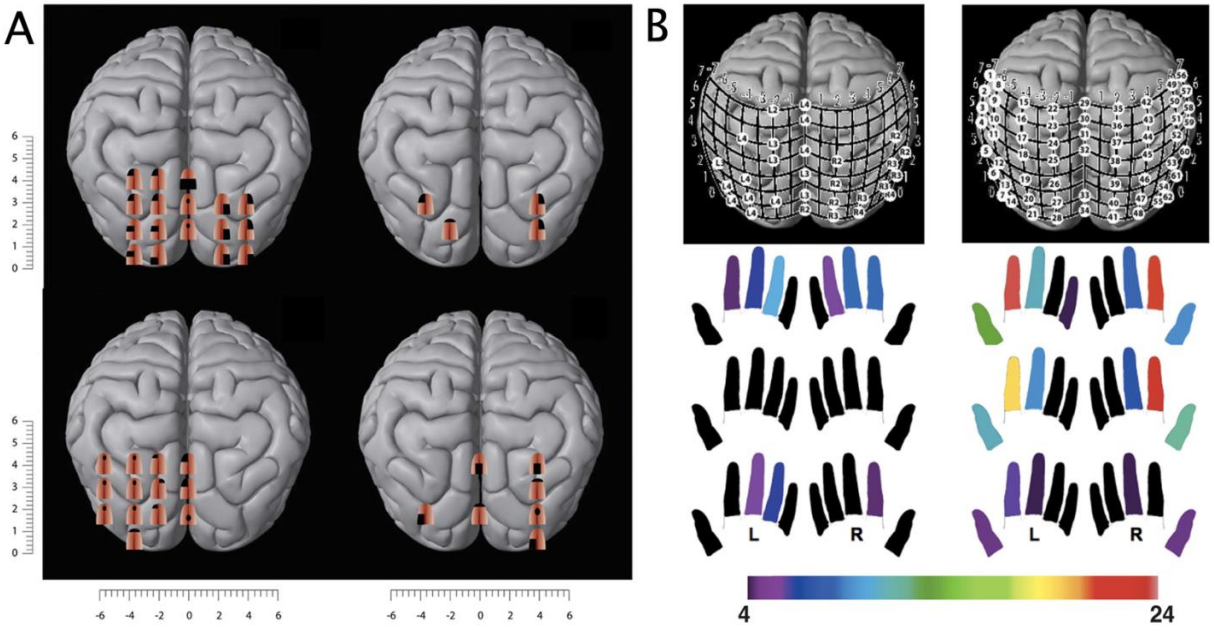


Figure I-24 Sensations tactiles induites chez un aveugle congénital par stimulation magnétique transcrânienne du cortex visuel. (a) Sensations tactiles sur la langue après une semaine d'utilisation du TDU. (b) Sensation tactiles sur les doigts chez deux liseurs de Braille expérimentés (le code couleur correspond au nombre de sites du cortex visuel dont la stimulation peut induire une paresthésie du doigt correspondant).

[1] La réalisation de tâches assez simples exige souvent des temps d'exécution pouvant être jusqu'à 20 fois plus longs que ceux de personnes voyantes.

Pascual-Leone et Torres ont par exemple montré l'augmentation des régions sensorimotrices associées à l'index droit en comparaison du gauche, après l'apprentissage du Braille [Hamilton and Pascual-Leone, 1998; Pascual-Leone and Torres, 1993], ainsi que l'activation du cortex visuel lors de sa lecture, comme illustré dans la Figure I-25 tirée de [Hamilton and Pascual-Leone, 1998].

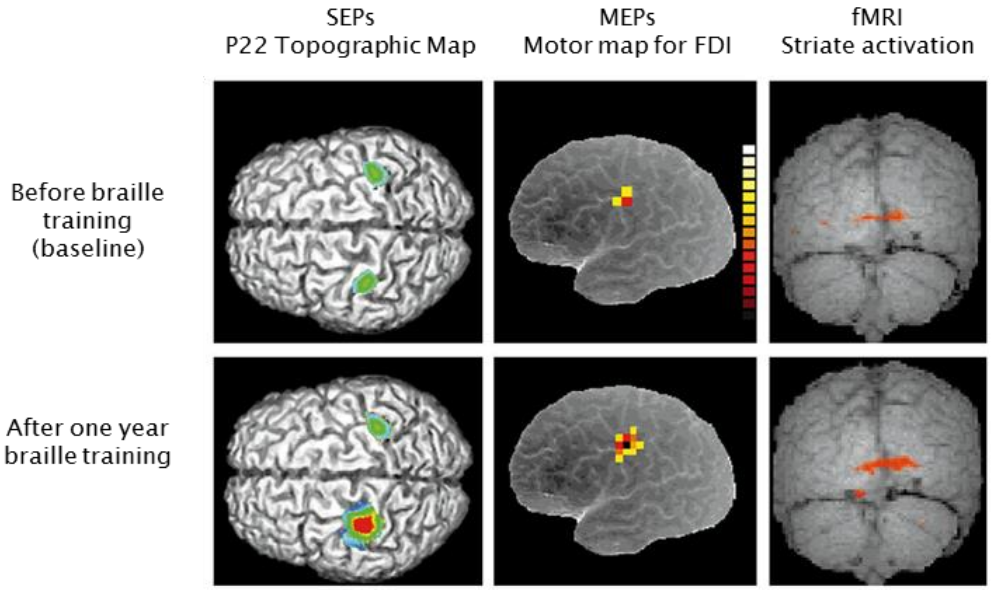


Figure I-25 Plasticité dans le cortex sensori-moteur et occipital suite à l'apprentissage du Braille par des sujets aveugles précoces, visible grâce à l'IRM fonctionnelle (fMRI), aux potentiels évoqués tactiles (SEPs) produits par stimulation mécanique du bout de l'index, et à la localisation par TMS des aires motrices impliquées dans l'activité du muscle interosseux dorsal permettant le mouvement latéral de l'index (Motor map for FDI).

Il a aussi été mis en évidence que la stimulation électrique de la langue chez les utilisateurs aveugles du TDU provoquait une activité dans le cortex visuel [Ptito and Kupers, 2005], allant même jusqu'à reproduire la ségrégation entre la forme et le mouvement observée chez les voyants dans les voies visuelles ventrale et dorsale [Ptito et al., 2009]. Différentes études ont également montré des similarités d'activation corticale entre voyants et non-voyants équipés du TDU, dans des tâches visuo-spatiales de navigation [Chebat, 2010; Ptito et al., 2005]. Ces résultats ne sont pas propres aux dispositifs électrotactiles, des activations du complexe occipital latéral ayant aussi été observées lors de l'utilisation de systèmes visuo-auditifs comme The vOICe ou PSAV (mais pas lors de l'écoute de sons naturels) [Amedi et al., 2007; Arno et al., 2001a].

Citons pour finir les résultats de Ptito et Kupers utilisant la stimulation magnétique transcrânienne (TMS) du cortex occipital. Celle-ci entraîne chez les voyants la perception de phosphènes visuels[1]. Les sujets utilisant le TDU ont eu, pour leur part, l'apparition de sensations sur la langue lors de cette même stimulation, et des lecteurs expérimentés de Braille, des sensations au niveau des doigts [Kupers et al., 2006; Ptito et al., 2008], tel qu'illustré dans la Figure I-24 tirée de [Kupers et al., 2011].

Tous ces travaux, ainsi que d'autres non mentionnés ici, attestent de façon sûre de la plasticité corticale et des réorganisations résultant de l'usage de systèmes de substitution sensorielle. Ceci explique en partie l'engouement scientifique pour ces méthodes, qui même lorsqu'elles ne se montrent pas concluantes dans une optique d'assistance aux non-voyants, présentent un grand intérêt pour la compréhension des mécanismes de perception et de restructuration fonctionnelles. Bach-Y-Rita lui-même reconnait cette approche dans [Bach-y-Rita, 1983] :

> *The Tactile Vision Substitution System program was conceived, to a large extent, as a model for studying brain plasticity.*

[1] Les phosphènes sont des points lumineux perçus dans le champ visuel.

2.2 Neuroprothèses

A l'inverse des systèmes de substitution sensorielle, cherchant de manière non-invasive à exploiter les mécanismes de plasticité cérébrale entre modalités sensorielles, l'implantation de neuroprothèses vise à stimuler directement les voies visuelles. L'image, toujours acquise par une caméra, peut être « envoyée » par une stimulation électrique sur la rétine [Chow et al., 2004; Grumet et al., 2000; Humayun et al., 1999; Javaheri et al., 2006; Zrenner et al., 1997], le nerf optique [Brelén et al., 2006; Oozeer et al., 2005; Veraart et al., 1998], les corps géniculés latéraux [Marg and Dierssen, 1965; Nashold Jr, 1970], ou directement sur la surface corticale [Dobelle, 2000; Fernández et al., 2005], comme illustré dans la Figure I-26.

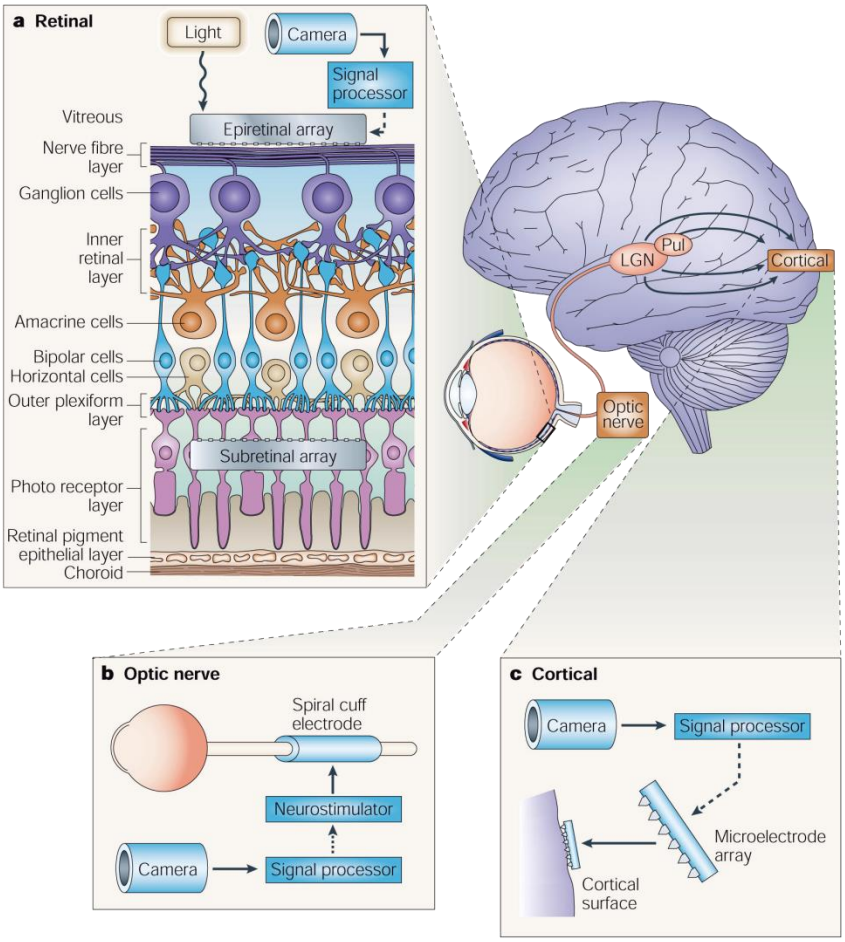


Figure I-26 Résumé des différentes localisations possibles d'implants visuels.

Ce type de dispositifs existe pour l'audition depuis la fin des années 60 et est maintenant très répandu [Niparko, 2009; Wilson and Dorman, 2008]. Appelés implants cochléaires, ils consistent en une partie externe, composée d'un microphone, d'un micro-processeur et d'une antenne. Les sons captés sont filtrés puis traités au niveau fréquentiel

pour sélectionner les électrodes à activer, puis ce signal est envoyé à la partie interne du dispositif, disposée sous la peau. Une fois acquis par l'antenne réceptrice, il est converti en impulsions électriques qui sont transmises aux électrodes implantées dans l'oreille interne venant stimuler la cochlée, qui à son tour acheminera ces informations au cortex via le nerf auditif (voir Figure I-27). La plupart des surdités totales, profondes ou sévères, récentes ou anciennes, peuvent désormais être réhabilitées par cette technologie dont plus de 200.000 personnes à travers le Monde sont équipées.

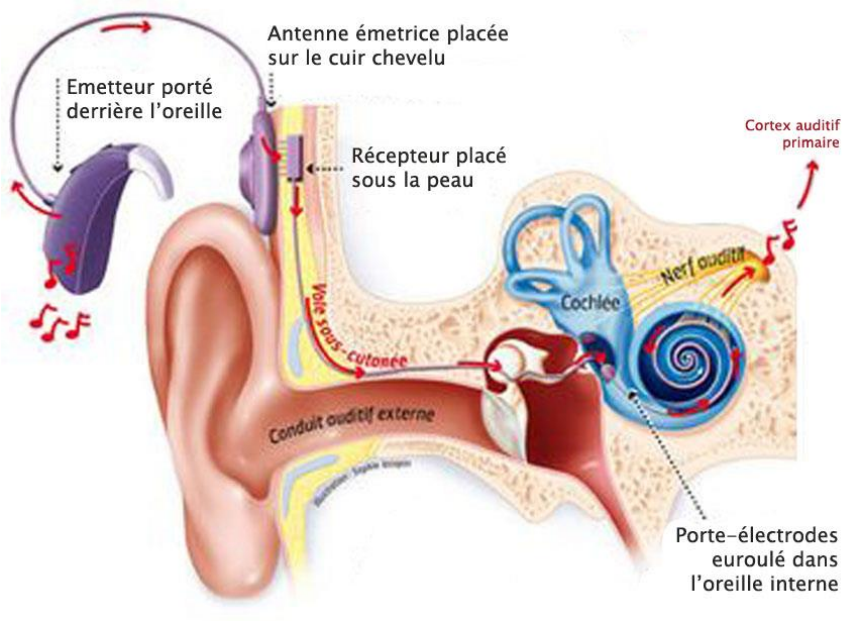


Figure I-27 Schéma d'un implant cochléaire.

Le cas des neuroprothèses visuelles reste cependant bien plus délicat. Les informations visuelles sont par nature beaucoup plus complexes et riches que celles transmises par des sons. Ceci se traduit notamment par le fait que, dans le cerveau humain, la capacité de transfert des signaux visuels est plus de 40 fois supérieure aux signaux sonores, si l'on se base sur le nombre respectif de fibres du nerf optique (environ 1,2 millions) et du nerf auditif (moins de 30.000). Les implants cochléaires utilisent en général 10 à 20 électrodes, qui s'avèrent suffisantes pour beaucoup de tâches, même aussi complexes que la compréhension du langage. La reconnaissance d'un visage, ou l'interprétation d'une scène en exigerait un nombre bien plus conséquent. C'est là le souci majeur auquel sont confrontés les projets visant au développement d'une neuroprothèse visuelle fonctionnelle.

L'origine des neuroprothèses visuelles remonte aux années 50-60, où plusieurs travaux ont montré qu'une stimulation électrique appliquée à la surface du cortex visuel, dans le lobe occipital, provoquait des sensations visuelles chez le patient opéré [Brindley and

Lewin, 1968; Penfield and Rasmussen, 1950]. Ces percepts, décrits par plusieurs sujets, s'apparentent à des points lumineux nommés phosphènes, de taille, de forme et de couleur variables, comme en témoigne la Figure I-28. En faisant varier la fréquence ou l'intensité de stimulation, il est possible d'évoquer des phosphènes plus ou moins lumineux. Néanmoins, en dehors de leur position qui dépend de la localisation de l'électrode, leurs autres propriétés (comme la couleur ou la taille), sont très dures à prévoir, qui plus est à contrôler. Si la stimulation d'une électrode permet la perception d'un point lumineux, alors il pourrait être possible de reproduire des motifs visuels grâce à une matrice d'électrodes, comme illustré dans la Figure I-28, tirée de [Normann, 2007].

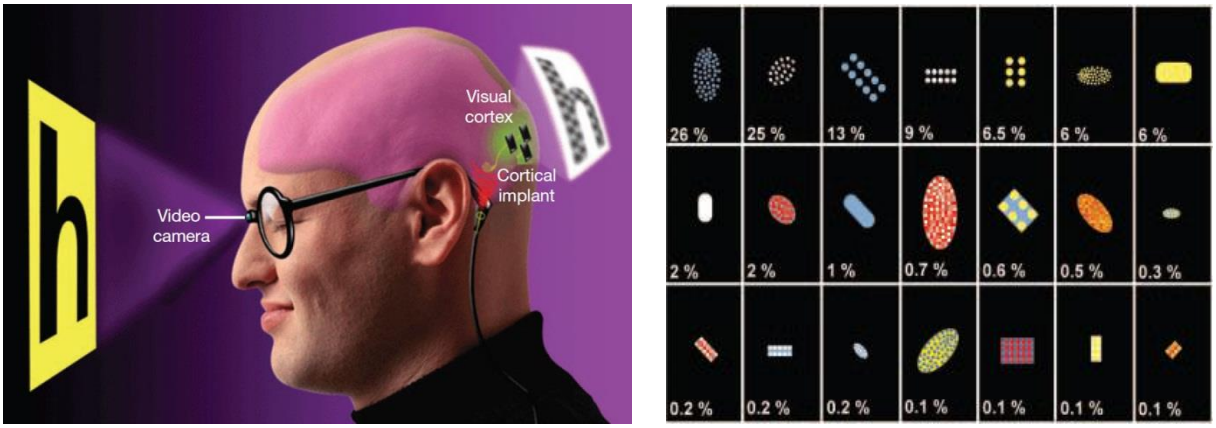


Figure I-28 A gauche : fonctionnement théorique d'une neuroprothèse corticale ; A droite : exemples de phosphènes perçus par un sujet implanté au niveau du nerf optique. Chaque dessin correspond à la description du patient, et le pourcentage indique sa fréquence d'occurrence. La taille représentée est arbitraire, les tailles perçues pouvant aller de 1 à 10° d'angle visuel.

Les prothèses visuelles reposent sur trois grands principes physiologiques : des courants électriques peuvent se substituer aux photons pour produire des percepts visuels (les phosphènes) ; la plupart des étiologies de cécité laissent les structures en aval intactes ; l'organisation rétinotopique des structures visées, conservée malgré la perte de la vision, permet une retranscription spatiale des images acquises par les caméras. Il s'ensuit que la plupart des implants visuels ont une architecture semblable, consistant en une caméra, un module de traitement visuel, un ensemble d'électrodes pour la stimulation, un système d'émetteur(s) et de récepteur(s) pour la transmission des données entre composants, et une source d'énergie pour le fonctionnement de la prothèse. Le module de traitement (qui peut être réparti sur les différents dispositifs, internes et externes) consiste généralement en la séquence décrite dans la Figure I-29 (adaptée de [Tsai et al., 2009]). Elle comprend l'acquisition de l'image, suivie de différents filtres visant à améliorer celle-ci (comme l'extraction des contours, ou le renforcement des contrastes), puis sa « pixellisation », pour correspondre à la taille de la matrice de stimulation, et enfin sa conversion en impulsions électriques.

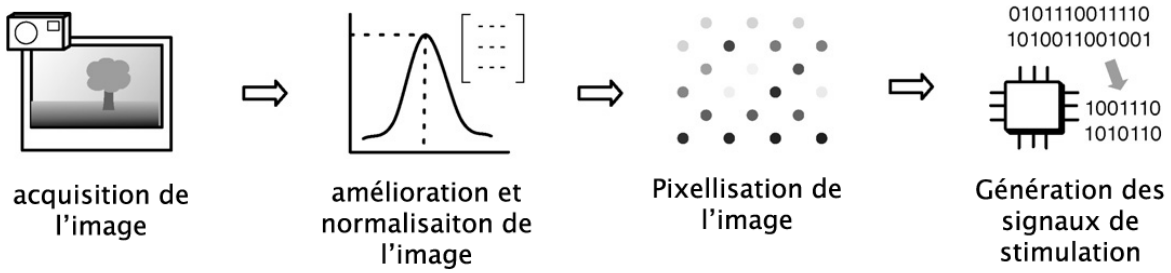


Figure I-29 Principe général de la chaine de traitement des neuroprothèses visuelles

Si certaines stimulations du corps géniculé latéral[1] (CGL), pour lesquelles les sujets étaient parfois capables de percevoir des points colorés, ont été reportées par Nashold ou Mara [Marg and Dierssen, 1965; Nashold Jr, 1970], les difficultés chirurgicales pour accéder à cette partie du cerveau ont largement entravé la poursuite de recherches dans cette voie. Nous nous intéresserons donc uniquement aux prothèses implantées dans la rétine, le nerf optique et le cortex visuel.

2.2.1 Implants corticaux

A la suite des travaux de Penfield et Rasmussen, ayant montré qu'une stimulation électrique du cortex visuel avec des électrodes de surface provoquait des percepts visuels [Penfield and Rasmussen, 1950], les premières prothèses corticales furent développées dans les années 70 par l'équipe de Brindley [Brindley and Lewin, 1968] et de Dobelle [Dobelle et al., 1979, 1974]. Aujourd'hui, de nombreux groupes ont développé ou continuent de mettre au point ce type de systèmes [Dobelle, 2000; Fernández et al., 2005; Normann et al., 2009, 1999; Schmidt et al., 1996; Troyk et al., 2003], mais peu de patients ont été implantés de façon chronique.

Les premiers implants testés sur des sujets humains, dont les larges électrodes étaient simplement posées à la surface du cortex, nécessitaient des courants de stimulation importants, provoquant de nombreux effets indésirables tels que des douleurs, des maux de tête, et même dans certains cas des crises d'épilepsie [Merabet et al., 2005; Normann et al., 2009]. De plus, les phosphènes évoqués étaient relativement larges, et leurs propriétés spatiales imprédictibles du fait des forts courants, conduisant à des non-linéarités dans les interactions entre électrodes lors d'activations multiples. Un deuxième type de prothèses dites intra-corticales a donc été mis au point, dont les microélectrodes (enfoncées dans le

[1] Le CGL est un relais synaptique du thalamus, recevant les afférences de la rétine et transmettant les informations visuelles par des projections vers le cortex visuel primaire au niveau de l'aire V1 du lobe occipital. Il est situé dans le diencéphale, juste au-dessus du tronc cérébral.

cortex) nécessitent moins de courant et offrent ainsi une stimulation plus précise [Bak et al., 1990; Bradley et al., 2005; Schmidt et al., 1996]. Ces prothèses, représentées dans la Figure I-30, arrivent seulement à leur stade d'essais cliniques, après de nombreux tests sur des animaux.

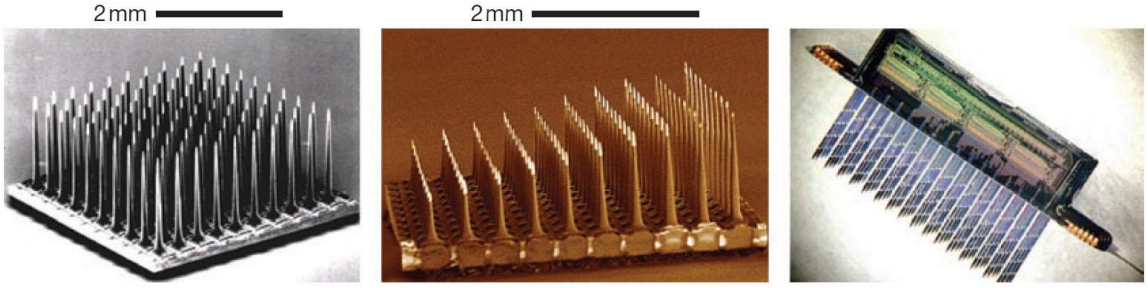


Figure I-30 Interfaces corticales. De gauche à droite : Utah Electrode Array (UEA) et Utah Slanted Electrode Array (USEA), contenant chacune 100 microélectrodes ; Michigan Three Dimensional Electrode Array

La chaîne de traitement habituelle des prothèses corticales consiste à appliquer aux images capturées par les caméras une série de filtres spatio-temporels qui reproduisent les fonctions des neurones de la rétine (à savoir les photorécepteurs, les cellules horizontales, bipolaires, amacrines et ganglionnaires), produisant une carte d'activation comme celle illustré dans la Figure I-31, qui est ensuite réduite pour correspondre à la taille de la matrice d'électrodes, puis convertie en une série d'impulsions pour chaque électrode, en fonction de leur position et de la taille de leur champ récepteur [Fernández et al., 2005].

Les implants corticaux présentent l'avantage de pouvoir être utilisés dans les cas où les structures de l'œil ou du nerf optique ont été trop endommagées suite à un traumatisme ou une maladie. Cependant, les traitements neuronaux sont d'une nature très complexe au niveau du cortex visuel, rendant extrêmement délicat le choix d'une stratégie d'encodage de l'image en stimulations électriques. Ces difficultés expliquent la direction qu'ont pris la plupart des travaux ultérieurs, cherchant à stimuler le système nerveux plus en amont des voies visuelles (au niveau de la rétine ou du nerf optique), où les réseaux de neurones et leurs traitements sont mieux connus et moins complexes que ceux du cortex et du thalamus.

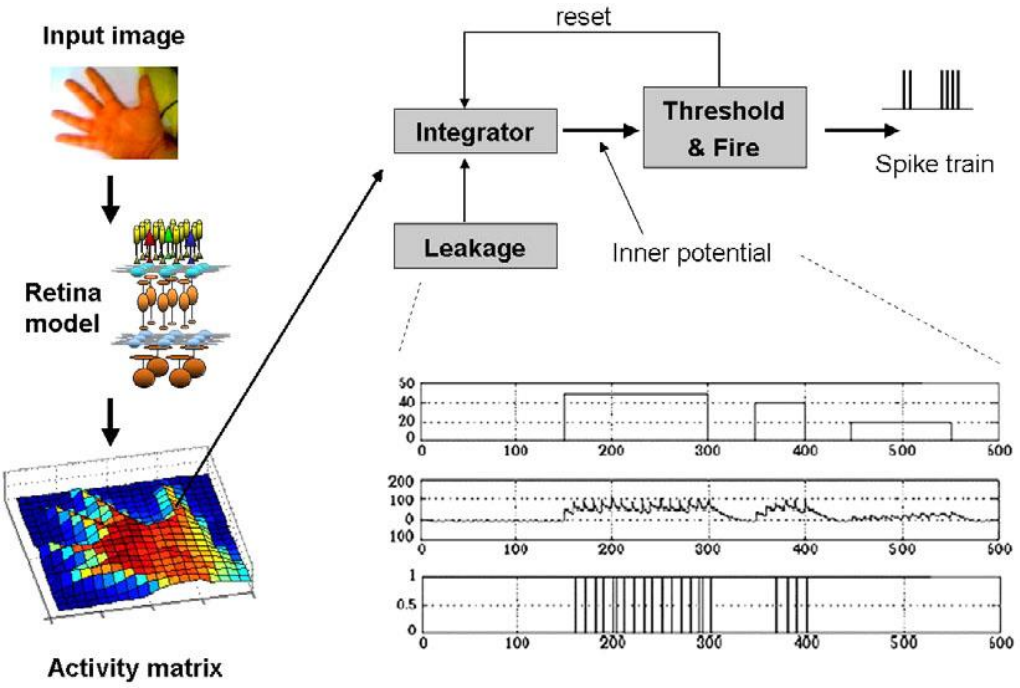


Figure I-31 Organisation d'un modèle de rétine pour prothèse visuelle corticale (tirée de [Fernández et al., 2005])

2.2.2 Implants rétiniens

Depuis les années 90, un nombre croissant de groupes de recherche se penche sur le développement de prothèses rétiniennes, consistant à stimuler les neurones encore fonctionnels de la rétine de patients atteints de pathologies dégénératives, telles que la rétinite pigmentaire (RP) ou la DMLA [Humayun et al., 1999]. En effet, plusieurs études morphométriques post-mortem effectuées sur des malades souffrant de RP n'ayant plus de perception visuelle avant le décès, ont montré que si moins de 4% des noyaux cellulaires de la couche externe de la rétine (contenant les photorécepteurs) subsistaient, ils étaient près de 30% dans la couche intermédiaire, et presque 80% dans la couche interne contenant les neurones ganglionnaires [Santos A et al., 1997; Stone JL et al., 1992]. Il est par conséquent possible d'envisager de stimuler ces neurones résiduels au niveau épi-rétinien [Eckmiller, 1997; Humayun et al., 2003; Rizzo and Wyatt, 1997] ou sous-rétinien [Chow et al., 2004; Zrenner et al., 1997].

Les systèmes utilisant un implant épi-rétinien reposent pour la plupart sur une architecture composée d'une caméra montée sur des lunettes, et d'une unité de traitement acquérant et convertissant l'image, ensuite envoyée par un canal de transmission sans-fil (laser ou radio) à une grille d'électrodes 2D implantée à l'intérieur de l'œil, juste au-dessus

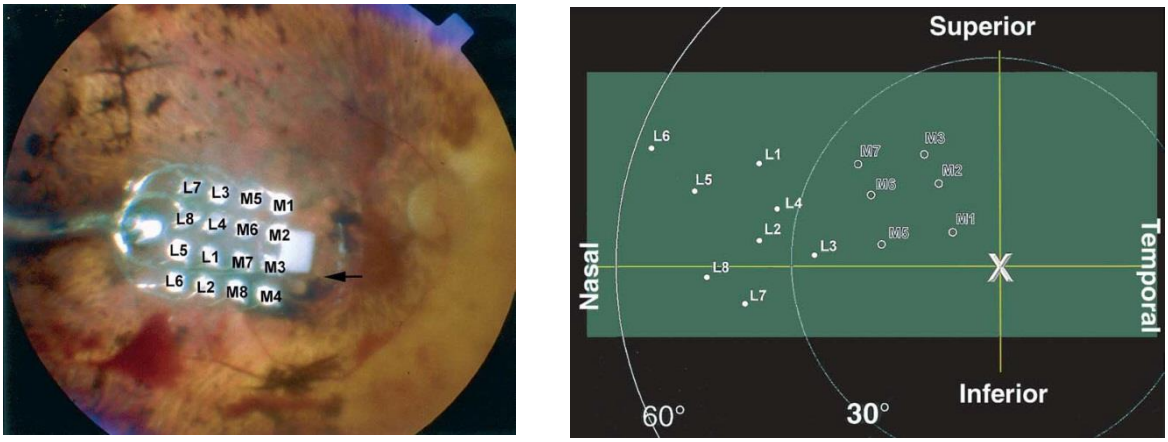


Figure I-32 Image de l'implant sous-rétinien développé par Humayun et al., et de la localisation dans le champ visuel des phosphènes associés à la stimulation de chacune des électrodes.

de la rétine. Chaque électrode, stimule les neurones rétiniens sous-jacents par de faibles courants électriques, provoquant la perception de points visuels dont la position dans le champ visuel varie selon le site de stimulation (voir Figure I-32). Les premiers tests ont montré que ces phosphènes permettaient de détecter des mouvements, la présence de lumière, ou encore de reconnaître des formes simples [Humayun et al., 2003]. Sur ce principe ont été développés depuis 2006 les systèmes Argus puis Argus II par la compagnie californienne Second Sigh[1]. L'argus II (voir Figure I-33), approuvé pour usage clinique et commercial dans l'union européenne en 2009, et aux Etats-Unis en 2013, a maintenant été implanté chez près de 70 personnes, malgré son coût d'environ 100 000 dollars. S'il possède 60 électrodes, l'acuité maximale relevée reste extrêmement faible, autour de 20/1260, ce qui est largement en dessous du seuil de cécité (20/500 selon le critère de l'OMS, 20/200 pour celui en vigueur aux Etats-Unis), mais permet la détection de formes simples et contrastées [Ahuja et al., 2010].

Le deuxième type de prothèses rétiniennes, les implants sous-rétiniens, n'utilise pour sa part aucune caméra. Ils sont composés de photodiodes (environ 5000 pour le prototype proposé par Chow [Chow et al., 2004], 1500 pour celui de Zrenner commercialisé par Retinal Implant [Rizzo et al., 2014]), chacune associée à une électrode de stimulation. La lumière capturée par ces diodes est convertie en signal électrique stimulant la couche voisine de neurones. Certains de ces implants sont dits passifs, car alimentés seulement par la lumière traversant la rétine, d'autres actifs, car ils utilisent une source d'énergie externe, la puissance de stimulation ayant souvent été jugée trop faible dans le premier cas. Une société allemande, Retina Implant GmbH, fondée par Zrenner, a mis au point un de ces implants sous-rétiniens de 40 par 40 photodiodes. Celles-ci sont reliées par un câble à une bobine

[1] http://2-sight.eu/fr/

disposée sous le crâne au niveau de l'oreille. Une batterie externe aimantée, portée derrière l'oreille, permet d'alimenter le système. Les premiers tests sur des patients montrent que les performances de ce dispositif restent, comme pour les implants épi-rétinien, relativement faibles, avec une acuité d'environ 20/1000.

Figure I-33 Implant sous rétinien Argus II, commercialisé par Second Sight

2.2.3 Implants dans le nerf optique

L'équipe de Claude Veraart, à l'Université de Louvain en Belgique, a développé un autre type d'implants, consistant en une électrode à manchon spiral [Delbeke et al., 2002; Veraart et al., 1998]. Le premier sujet, âgé de 59 ans, fut implanté en 1998, au niveau du segment intracrânien du nerf optique [Veraart et al., 1998]. Les stimulations électriques ont pu induire différents phosphènes, dont les tailles et positions sont reportées dans la Figure I-34, qui permettent de reconnaître certaines formes simples [Brelén et al., 2005; Delbeke et al., 2003; Veraart et al., 2003]. Une nouvelle approche, développée par la même équipe, consistait à positionner les électrodes sur la partie antérieure du nerf optique, au niveau intra-orbital. Cette technique permet une chirurgie plus courte et présente moins de risques que la première, qui nécessite une craniotomie [Brelén et al., 2006; Oozeer et al., 2005]. Un second patient, de 68 ans, a pu être opéré en 2006 en utilisant cette méthode. Plus récemment, le projet C-Sight (Chinese project for Sight), a mis en place une technique

d'implantation similaire, grâce à des électrodes pénétrant le nerf optique, pour le moment testée uniquement sur des animaux [Sui et al., 2009].

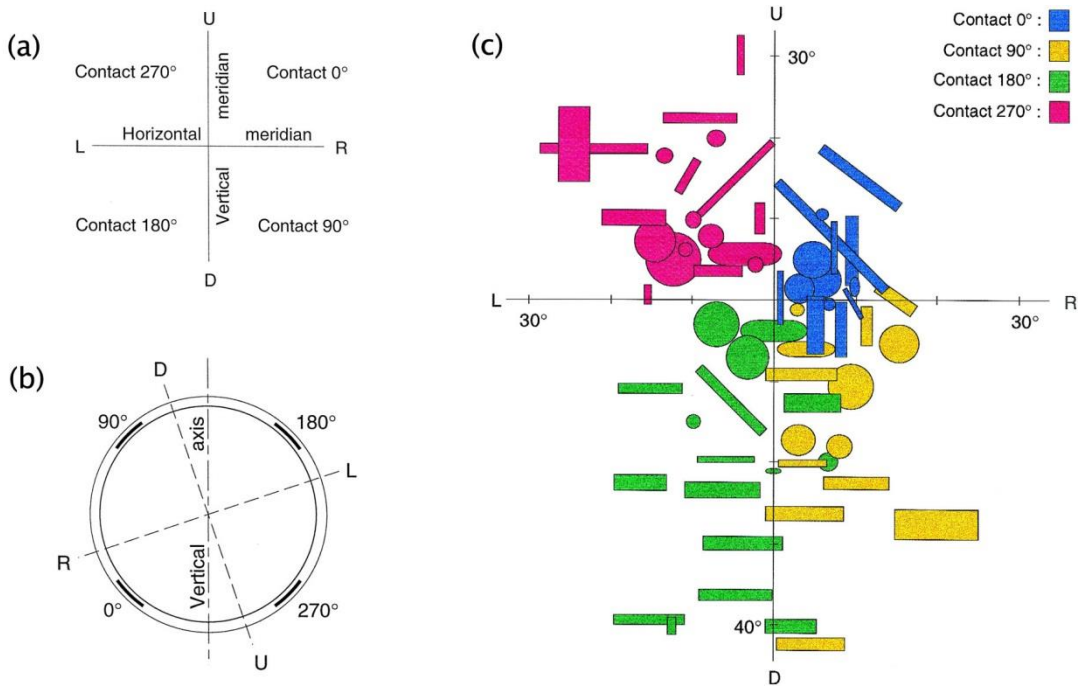


Figure I-34 Organisation rétinotopique du nerf optique. (a) Position des 4 contacts autour du nerf ; (b) Relation théorique entre la position des contacts et la position des phosphènes dans le champ visuel ; (c) Phosphènes décrits par le patient implanté (les couleurs correspondent au contact utilisé, pas au percept)

D'autres évaluations du dispositif de Veraart ont montré qu'en plus de reconnaître des motifs simples, il pouvait permettre de localiser, d'identifier et de saisir des objets [Duret et al., 2006]. L'expérience a été réalisée en plaçant un sujet face à une table noire divisée en 9 secteurs (voir Figure I-35), sur laquelle était disposé un des 6 objets connus du sujet (une grande et une petite bouteille, un boîtier de CD, une tasse, un couteau et un tube de dentifrice). Le sujet devait localiser l'objet par des balayages visuels de la table. Puis, une fois ce dernier trouvé, il devait effectuer d'autres scans verticaux et horizontaux afin de l'identifier et de l'attraper. Après plusieurs sessions, le sujet était capable de réaliser ces trois tâches avec près de 100 % de réussite. En moyenne, 30 secondes étaient nécessaires pour la localisation, 40 à 50 pour l'identification, et 4 ou 5 pour la préhension. Même si ces tests semblent concluants, il est important de souligner que les objets étaient peints en blanc, posés sur un fond noir, très proches du sujet, et toujours orientés de la même façon. Ces résultats ne peuvent donc pas se transposer à une situation réelle, il est par conséquent abusif de considérer que le sujet est capable de localiser et d'identifier les objets de façon générale.

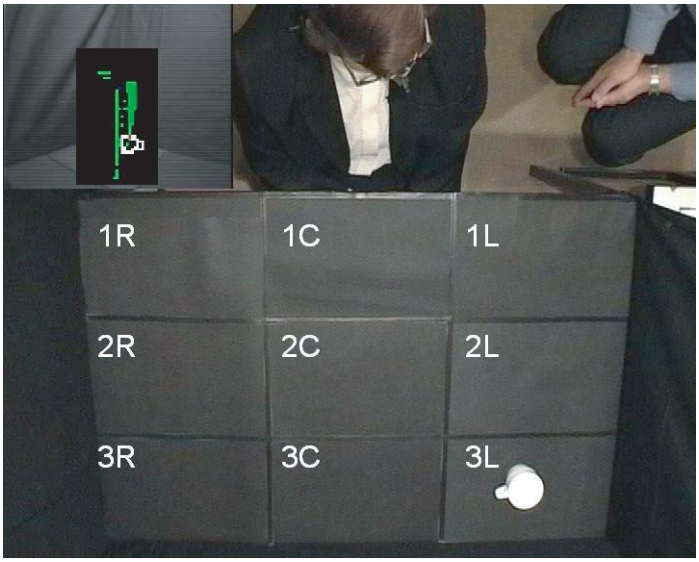


Figure I-35 Dispositif expérimental de test de l'implant du groupe de Claude Veraart. Le sujet est placé face à une table divisée en 9 zones, sur une desquelles est placé un objet devant être localisé, identifié et saisi.

2.3 Conclusion sur l'approche holistique

Les neuroprothèses, que nous venons de décrire, peuvent offrir un espoir de restaurer certaines fonctions visuelles à moyen ou long terme, mais les dispositifs actuels, rarement implantés de façon durable, souffrent toujours de nombreuses limitations. Il s'agit en effet de méthodes invasives, nécessitant des actes chirurgicaux relativement lourds qui présentent des risques opératoires. De plus, le coût financier de ces dispositifs est à l'heure actuelle extrêmement élevé (autour de 100 000 dollars pour l'Argus II). Beaucoup de ces projets sont encore au stade d'essais cliniques et n'ont donc pu être implantés de façon durable. Ceux qui l'ont été présentent des résolutions encore trop faibles pour être réellement utilisables au quotidien (une soixantaine d'électrodes pour les implants rétiniens, une centaine pour les corticaux). L'acuité visuelle procurée reste en effet très faible, autour de 20/1000 pour la plupart[1]. La miniaturisation des matrices de stimulation pourra certes permettre une relative augmentation de la résolution, mais les interférences existant entre des électrodes trop proches semblent limiter le développement de prothèses haute-résolution [Dagnelie, 2008]. Un certain nombre d'autres problèmes, inhérents à chaque type

[1] Un sujet non-voyant équipé de la prothèse corticale de l'institut Dobelle présentait par exemple une acuité de seulement 20/1200, malgré un long entraînement et plusieurs années d'implantation.

Type de cécité	Méthodes de réhabilitation visuelle viables			
	Prothèses pré-chiasmatiques	*Prothèses corticales*	*Substitution sensorielle*	*Aides visuelles*
Tardive (rétine touchée)	✓	✓	✓	✓
Tardive (structure périphérique touchée)		✓	✓	✓
Précoce (pendant le développement)			✓	✓
Précoce (après le développement)			✓	✓

Tableau I-2 Méthodes de réhabilitation en fonction du type de cécité (figure adaptée de [Veraart et al., 2004]).

d'implant selon sa localisation, freinent leur développement, comme la biocompatibilité, ou les questions d'alimentation et de transmission [Cohen, 2007]. Enfin, ces solutions restent restreintes à certaines pathologies, en particulier celles de la rétine ou du nerf optique. Car si les implants corticaux pourraient, eux, s'appliquer à un plus grand nombre, aucune neuroprothèse n'a à ce jour été implantée à des aveugles de naissance. Il semble en effet nécessaire que le cortex visuel ait d'abord pu se développer de façon normale pour pouvoir évoquer des percepts lors d'une stimulation électrique [Veraart et al., 2004]. La stimulation transcrânienne (TMS) de V1 n'entraîne par exemple aucun phosphène chez des aveugles précoces et un nombre important d'experts s'accordent à dire qu'au-delà de la phase critique de développement, les possibilités de réhabilitation par neuroprothèses sont très faibles [Kay, 1984; Merabet et al., 2005; Veraart et al., 2004]. Cette question fait néanmoins débat, notamment au vu des résultats obtenus avec les implants cochléaires, ayant permis à des sourds congénitaux de recouvrer des fonctions auditives. De façon similaire, l'implantation d'une neuroprothèse visuelle dans la petite enfance pourrait permettre de regagner certaines fonctions visuelles grâce à des mécanismes de plasticité, mais à l'heure actuelle, étant donnée la résolution des implants, cette possibilité semble incertaine, tel que le souligne par exemple Gislin Dagnelie [Dagnelie, 2008] :

> *There is broad consensus that functional vision restoration is predicated on prior visual experience. Although it is true that cochlear implants have now proven successful in bringing auditory function to congenitally deaf children implanted within the first year of life, it is not clear that visual prostheses will have the same benefit when implanted in early childhood: The intricate development potential of the*

normal visual system that starts at birth and extends over approximately a decade cannot be realized through an input device with a few hundred channels.

A l'inverse des neuroprothèses, les systèmes de substitution sensorielle sont adaptés à tout type de cécité ou de malvoyance, et présentent l'avantage d'être non-invasifs et généralement peu coûteux. Comme nous l'avons vu dans la section I.2.1, les études effectuées avec les dispositifs auditifs ont montré la possibilité de reconnaissance et de localisation de formes simples ainsi que de formes plus complexes comme des objets de la vie courante [Arno et al., 2001b, 1999; Auvray, 2004; Auvray et al., 2007; Cronly-Dillon et al., 2000, 1999; Hanneton et al., 2010]. Après plusieurs années d'utilisation de The vOICe, certains sujets aveugles rapportent même une véritable expérience visuelle, de nature automatique, qui s'affranchit du mode de restitution utilisé [Ward and Meijer, 2010].

Des résultats similaires ont été observés avec des systèmes visuo-tactiles [Bach-y-Rita et al., 1998; Jansson, 1983; Kaczmarek et al., 1997; Kaczmarek and Haase, 2003; Sampaio et al., 2001]. Il a été montré qu'ils pouvaient être utilisés pour la lecture (du moins pour la reconnaissance de caractères) dans [Craig, 1981; Loomis, 1974], et aussi qu'ils permettaient d'effectuer des jugements perceptifs tels que la perspective, la parallaxe, les ombres, l'interposition des objets, ou des estimations concernant la profondeur [Bach-y-Rita et al., 1969a; Epstein, 1985].

Si ces résultats pourraient sembler prometteurs, l'acuité visuelle résultant de l'utilisation de ces systèmes reste cependant très faible. Différentes études ont évalué celle-ci en utilisant des adaptations des tests de Snellen. Pour des sujets équipés du Tongue Display Unit, l'acuité moyenne était de 20/430 [Chebat et al., 2007; Sampaio et al., 2001], alors que celles de 9 utilisateurs de The vOICe se situaient entre 20/200 et 20/600 après 50 à 100 heures d'apprentissage [Striem-Amit et al., 2012]. Bien que supérieures à celles relevées avec la plupart des neuroprothèses, elles restent malgré tout inférieures au seuil d'acuité visuelle relevant de la cécité. Le gain offert par ces différentes solutions reste par conséquent très modéré et leur efficacité s'en trouve relativement limitée.

En plus de montrer des performances insuffisantes pour une utilisation dans la vie courante, dues aux trop faibles résolutions, les dispositifs de substitution sensorielle souffrent de plusieurs autres inconvénients liés à leur modalité de restitution. Dans les cas visuo-auditifs, les sons produits interfèrent avec les sons ambiants, pouvant rendre difficile leur interprétation pour la navigation. Ils gênent également la communication orale, et peuvent aussi affecter l'équilibre [Velázquez, 2010]. Les systèmes tactiles quant à eux, s'ils sont moins touchés par cette « surcharge » sensorielle, présentent des effets indésirables comme des irritations, voire des douleurs, ou la contraction involontaire de muscles [Capelle et al., 1998; Kaczmarek et al., 1991; Szeto and Saunders, 1982].

En conclusion, si certains utilisateurs de dispositifs de substitution sensorielle ou de neuroprothèses peuvent apprendre à reconnaître des motifs visuels simples et éventuellement se déplacer dans l'espace, ou se saisir d'objets, ces systèmes n'ont pour la plupart été évalués que dans des tâches très simples en laboratoire. Ils n'ont par exemple été que très rarement évalués dans des tâches de navigation [Durette et al., 2008; Gomez Valencia, 2014] ou dans des contextes écologiques, où la plupart des tâches visuo-motrices s'avèrent bien plus complexes qu'une réponse à choix forcé entre une ligne verticale et horizontale. Leur utilisation apparaît donc difficilement transposable à la reconnaissance d'objets dans des environnements non-contrôlés, d'autant que la variabilité des performances entre sujets reste très importante. Presqu'aucun groupe de recherche n'a par exemple réussi à obtenir de bonnes performances dans une série de tâches visuelles sur l'ensemble (ou la majorité) des sujets testés [Bullier, 2002].

Il semble donc admis qu'aucun de ces systèmes ne peut être utilisé comme dispositif de suppléance dans des environnements naturels pour répondre aux besoins quotidiens des non-voyants. Une autre démarche consisterait à développer des aides spécifiques à une fonction donnée et donc utilisables pour pallier un besoin précis. Nous allons par exemple développer dans la section suivante, le cas des aides à la mobilité et à l'orientation -qui sont parmi les problématiques majeures du handicap visuel-, ainsi que les dispositifs permettant la reconnaissance et la localisation d'objets.

3. Systèmes d'assistance basés sur une approche fonctionnelle

Parmi les principales difficultés rencontrées au quotidien par les non-voyants arrivent en tête l'accessibilité à l'information, en particulier écrite, et la mobilité [Jacobson and Kitchin, 1997]. De nombreux dispositifs ont été mis au point afin de répondre aux attentes concernant l'accès à l'information, utilisant par exemple la synthèse vocale, la reconnaissance de caractères, l'alphabet braille ou encore les liseuses d'écrans. Ces systèmes s'avèrent être des solutions relativement efficaces, adoptées par une grande partie de la population non-voyante. Nous ne détaillerons pas plus ces aides ici mais nous nous intéresserons plutôt aux questions de la mobilité et de la navigation (nous reviendrons sur une définition plus précise de ces termes), ainsi qu'à la reconnaissance d'objets.

3.1 Aides à la navigation

La mobilité est un des facteurs majeurs contribuant à la qualité de vie. Elle est nécessaire à l'accomplissement d'un grand nombre d'activités, aussi bien dans les domaines sociaux, personnels que professionnels. De par la composante visuelle d'un grand nombre de comportements associés aux déplacements, les déficiences visuelles ont des conséquences majeures sur la navigation. Les problèmes relatifs aux déplacements constituent en effet la deuxième cause la plus importante de handicap ressenti par les non-voyants [Golledge, 1993]. Un certain nombre d'enquêtes ont permis d'identifier et de préciser ces besoins. Ainsi, dans une étude réalisée en Angleterre sur une population de jeunes non-voyants, il apparaît que 20 % des personnes interrogées n'ont pas quitté leur domicile au cours de la semaine précédente, 34 % ne sont sorties que dans le voisinage, et seulement 41 % se sont déplacées seules et à pied [Bruce et al., 1991]. Ces observations concordent avec les chiffres relevés par Clark-Carter, qui montrent que 30 % des non-voyants ne sont pas en mesure de naviguer seuls en dehors de leur résidence, et que parmi ceux qui s'y aventurent, la majeure partie tend à suivre des routes connues car l'exploration de lieux non familiers résulte souvent en une désorientation entraînant du stress, de la peur et de l'anxiété [Clark-Carter et al., 1986].

L'enquête Handicap-Incapacité-Dépendances conduite en France dans les années 2000 relève également ces difficultés [Bournot et al., 2005]. En effet, plus d'un déficient visuel sur deux (56 %) déclare une incapacité sévère concernant la mobilité et les déplacements (effectuer ses achats, sortir du domicile, porter des objets, monter ou descendre un étage d'escalier). Ce sont les déplacements à l'extérieur qui sont les plus

problématiques. Parmi les déficients visuels âgés de plus de 20 ans et physiquement aptes à se déplacer, 58 % éprouvent des difficultés dans ce type de situations. Près de 30 % des personnes interrogées sont incapables de se déplacer seules, 15 % uniquement sur certains itinéraires, et seules 14 % déclarent être en mesure de se déplacer dans de nouveaux lieux malgré les difficultés rencontrées [Sander et al., 2005]. Ces difficultés touchent d'autant plus les aveugles et malvoyants profonds : plus de 9 sur 10 rapportent des problèmes de mobilité et d'orientation, dont 63 % ne peuvent se déplacer seuls. Enfin, l'âge est un facteur aggravant, car le vieillissement s'accompagne souvent d'autres déficiences telles que des troubles moteurs, auditifs, ou mnésiques.

Avant d'aborder la perception de l'espace par les déficients visuels, les stratégies qu'ils adoptent pour se déplacer, et les systèmes qui ont été développés pour les aider dans cette tâche, il convient de préciser certaines notions. La notion de navigation, tout d'abord, se réfère à l'acte de déplacement vers une destination. Elle est souvent définie comme le processus permettant de se rendre d'un endroit à un autre, et englobe donc de nombreuses tâches, comme le fait d'identifier sa position, de choisir une route, de suivre un itinéraire ou encore d'éviter des obstacles. Elle implique par conséquent de multiples fonctions d'ordre perceptif, moteur et cognitif.

On distingue généralement deux principales composantes dans la navigation : l'orientation et la mobilité, aussi appelées navigation globale et navigation fine [Adams and Beaton, 2000; Guth and Rieser, 1997]. La première se rapporte à la faculté à se situer dans l'environnement (identifier sa position et son orientation relativement à une référence extérieure), à localiser les points de départ et d'arrivée, et à choisir un trajet à emprunter parmi les différentes routes possibles. La mobilité en revanche concerne les aspects plus immédiats, à savoir les comportements d'évitement d'obstacles et la faculté de garder un cap. Elle relève de la notion de locomotion.

Du point de vue de la cognition spatiale, la navigation met en jeu la notion de « wayfinding », qui implique notamment la résolution de problèmes spatiaux [Boumenir, 2011; Caddeo et al., 2006; Golledge et al., 1998; Ross and Blasch, 2000]. Ce processus se décompose en trois étapes :

1) La construction d'une carte cognitive : par l'exploration active ou passive (la description de l'environnement par un tiers par exemple), l'individu est capable de développer une représentation mentale d'un lieu. Cette information spatiale est encodée et gardée en mémoire pour pouvoir être exploitée par la suite.
2) La prise de décision : à partir de la carte cognitive de l'environnement, la prise de décision consiste à sélectionner le meilleur chemin vers son but, en fonction de différents facteurs comme la distance ou le temps de parcours, les obstacles

rencontrés, la notion de plaisir et de sécurité. Ce type de processus implique la résolution de problèmes de nature spatiale.

3) Exécution des décisions : consiste à mettre en application les raisonnements effectués précédemment. Il faut donc être capable de se souvenir de l'objectif et des différentes sections de l'itinéraire, ainsi que de conserver son orientation par rapport au trajet et à l'environnement.

Dans les processus d'acquisition d'informations spatiales, la vision apparaît comme la modalité sensorielle la plus impliquée [Foulke, 1982]. Une des composantes importantes de la navigation consiste en effet à anticiper et prétraiter la configuration spatiale de l'environnement, afin d'effectuer des choix lors de la traversée d'espaces complexes. Ces projections sont généralement possibles grâce au sens visuel, qui renseigne sur les différents éléments présents dans l'environnement et sur leur position. Avec une déficience visuelle, il devient par conséquent beaucoup plus difficile d'effectuer ce type de projections, les informations disponibles étant beaucoup plus éparses, car limitées à l'audition et au sens tactile, parfois aussi à l'odorat [Golledge, 1993]. Or ce type de perceptions contraint souvent à une exploration séquentielle de l'environnement, plus lente et plus coûteuse, qui se limite majoritairement à l'environnement immédiat autour du piéton. Au niveau cognitif, il s'ensuit donc que les représentations mentales construites s'apparentent à des routes plutôt qu'à des cartes, les informations spatiales étant acquises de façon égocentrée plutôt qu'allocentrée.

Les besoins, aussi bien en termes d'orientation que de mobilité sont donc nombreux et relativement bien identifiés. Plusieurs enquêtes ont montré par ailleurs qu'une grande partie des non-voyants est disposée à utiliser des technologies d'assistance, si elles permettent d'accroître leur autonomie, à condition qu'elles respectent certains critères de coût, d'efficacité, d'ergonomie, et d'esthétique. Pour répondre à ces attentes, un grand nombre d'aides électroniques portables ont vu le jour depuis une cinquantaine d'années (voir par exemple les dispositifs illustrés dans la Figure I-36).

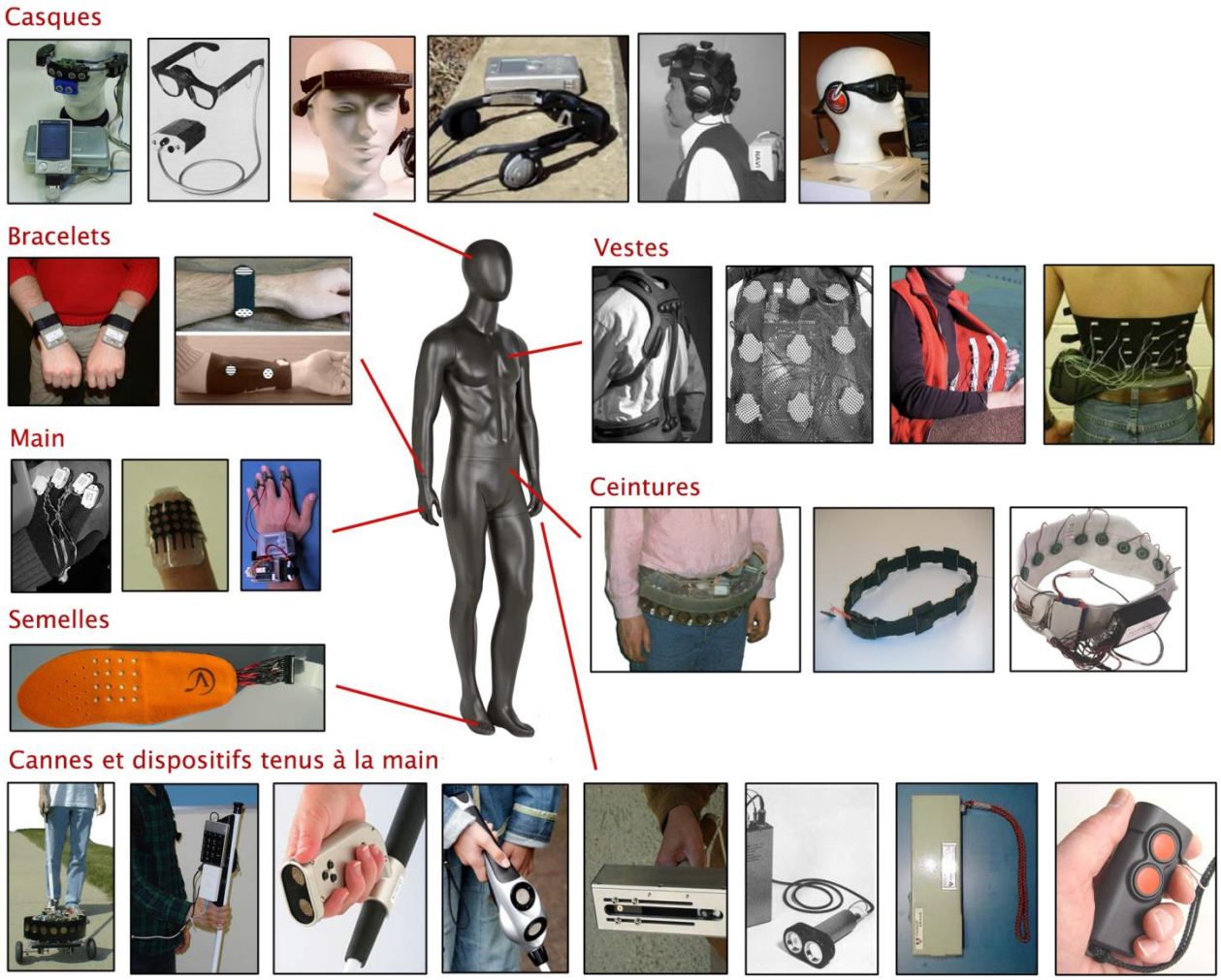


Figure I-36 Dispositifs portables d'aide aux non-voyants (les exemples de chaque catégorie sont données de gauche à droite) : casques (Florida University Portable 3D Sound/Sonar Navigation System [Aguerrevere et al., 2004], SonicGuide [Kay, 1974], Kaspa, Virtual Acoustic Space [Gonzalez-Mora et al., 1999], NAVI [Sainarayanan et al., 2007], Tyflos [Dakopoulos and Bourbakis, 2008; Dakopoulos et al., 2007]) ; vestes (Kahru Tactile Outdoor Navigator [Gemperle et al., 2001], Shoulder-Tapping [Ross and Blasch, 2000], TNO Human Factors Vest [Van Veen and Van Erp, 2003], MIT Tactile Display [Jones et al., 2006]) ; ceintures (NavBelt [Shoval et al., 2003, 1998], ActiveBelt [Tsukada and Yasumura, 2004], FeelSpace Belt [Nagel et al., 2005]) ; bracelets (GentleGuide [Bosman et al., 2003] et prototypes de [Ng et al., 2007]) ; dispositifs portés sur la main (Guelph Tactile Glove [Zelek et al., 1999], Soft-actuator-based Tactile Display [Koo et al., 2008], FingerBraille [Amemiya et al., 2004]) ; semelles tactiles [Velazquez et al., 2009] ; cannes et dispositifs tenus à la main (GuideCane [Ulrich and Borenstein, 2001], KinectCane [Takizawa et al., 2012], BAT K-Sonar ["BayAdvancedTechnologies," n.d.], UltraCane ["UltraCane," n.d.], Virtual White Cane [Yuan and Manduchi, 2005], Kay Sonic Torch [Kay, 1964], Mowat Sensor [Pressey, 1977], Miniguide ["GDP Research," n.d.]).

3.1.1 Aides à la mobilité

Les aides traditionnelles à la mobilité sont la canne blanche et le chien guide. Si ces deux méthodes présentent de nombreux avantages, et ont permis d'accroître grandement l'autonomie de certains aveugles, elles présentent néanmoins certaines limitations et n'ont paradoxalement été adoptées que par une faible partie de la population non-voyante. En France on estime que seulement 26 % des aveugles ont recours à une de ces deux techniques [Sander et al., 2005], et aux Etats-Unis ils sont 10 000 à utiliser un chien guide, et 109 000 une canne blanche parmi les 1,1 millions de non-voyants. La préférence pour la canne blanche s'explique surtout par le coût d'un chien guide, environ 15000 € (auxquels s'ajoutent des frais annuels pouvaient s'élever jusqu'à 6000 €). La canne à l'inverse, ne coûte qu'une cinquantaine d'euros, mais nécessite un long apprentissage auprès d'instructeurs en mobilité et offre des possibilités plus limitées que celles du chien, dont les utilisateurs ont en général des vitesses de marche plus rapides. Son usage se restreint par exemple à une portée de 1 à 2 mètres face à l'utilisateur, et ne permet pas d'éviter les obstacles aériens ou des trous au niveau du sol.

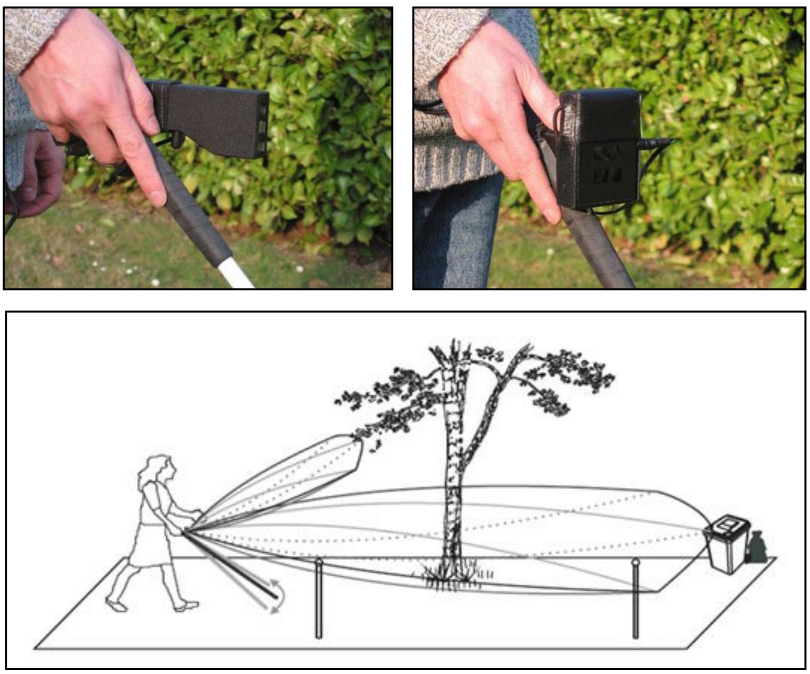


Figure I-37 Aides électroniques à la mobilité. En haut à gauche le système Teletact, à droite le Tom Pouce. En bas, illustration du fonctionnement de ces dispositifs permettant d'étendre la portée de la canne à 6 mètres, et de couvrir l'espace au niveau de la tête de l'utilisateur.

De nombreux dispositifs ont donc vu le jour depuis les années 60 afin d'augmenter la portée de la canne blanche par l'usage de systèmes laser ou acoustiques permettant un balayage plus large et plus lointain de l'environnement [Aguerrevere et al., 2004; Bissitt and Heyes, 1980; Bolgiano and Meeks, E., 1967; Kay, 1974; Takizawa et al., 2012; Ulrich and Borenstein, 2001]. Ceux-ci, portés sur un casque, fixés à la canne, ou tenus à la main, renvoient les informations sous forme tactile ou auditive à l'utilisateur. Parmi les plus populaires, citons le Tom Pouce [Damaschini et al., 2005] et le Teletact [René Farcy et al., 2006], illustrés dans la Figure I-37, ainsi que l'UltraCane [Hoyle and Dodds, 2006]. Pour une liste plus complète se reporter par exemple à la revue de la littérature de Roentgen [Roentgen et al., 2008] ou à [Dakopoulos and Bourbakis, 2010; Liu et al., 2010]. D'autres systèmes, visant également à éviter les obstacles sur le chemin de l'utilisateur, sont basés sur l'utilisation de caméras plutôt que de laser ou d'ultrasons. Ceux-ci seront présentés dans la section 3.2.1 de ce chapitre.

3.1.2 Aides à l'orientation

Si les aides électroniques à la mobilité facilitent les déplacements en permettant d'éviter les obstacles plus efficacement que les aides traditionnelles, elles ne fournissent cependant aucune information sur la localisation et l'orientation de l'utilisateur, permettant la planification d'itinéraire et à la compréhension spatiale de l'espace. Ces fonctions sont assurées par un autre type de dispositifs, appelé aides électroniques à l'orientation, ou EOAs[1].

Pour fournir à l'utilisateur des informations sur sa position, la direction à prendre, ou les points de repère qui l'entourent, le concept le plus répandu est l'utilisation de systèmes basés sur un capteur GPS, permettant de localiser le piéton dans un système de coordonnées absolues, couplé à des données cartographiques contenues dans ce que l'on nomme un Système d'Information Géographique (SIG), ainsi qu'à un moteur de calcul et de suivi d'itinéraire. Les instructions de guidage, ainsi que d'autres informations extraites du SIG sur les éléments environnants, sont généralement retournées à l'utilisateur par synthèse vocale, parfois également par des sons spatialisés ou par des dispositifs tactiles (comme des modules vibrants, positionnés à la taille, dans une veste ou aux poignets, mais aussi grâce à une tablette braille [Henze et al., 2006; Ross and Blasch, 2000; Tsukada and Yasumura, 2004]).

[1] Pour *Electronic Orientation Aid*.

Les premiers travaux ayant développé ce type d'architecture pour le guidage des non-voyants remontent aux années 80 [Brusnighan et al., 1989; Golledge et al., 1998, p. 198; Loomis, 1985]. Les équipes de Loomis et Golledge ont continué d'exploiter ces thématiques par la suite, et ont fourni de nombreux résultats sur la perception spatiale des déficients visuels et le développement d'aides à l'orientation adaptées [Golledge et al., 2004, 1998, 1995; Loomis et al., 2005, 2001, 1994]. A partir des années 90, de nombreux systèmes similaires ont vu le jour, comme le projet MoBIC [Petrie et al., 1997, 1996], Drishti [Helal et al., 2001; Ran et al., 2004], Strider [LaPierre, 1998], SWAN [Walker and Lindsay, 2006; B. N. Walker and Lindsay, 2005; Wilson et al., 2007] ou encore Odilia [Mayerhofer et al., 2008] parmi d'autres.

Figure I-38 Personal Guidance System (PGS) développé par Loomis et son équipe de l'Université de Californie (photographies du prototype original mis au point en 1993).

Indépendamment de ces projets de recherche, plusieurs EOAs ont été commercialisés et certains sont aujourd'hui utilisés par une partie de la population non-voyante. Parmi ces derniers, nous pouvons mentionner le Trekker, commercialisé par Humanware, qui fournit vocalement les noms des rues et les points d'intérêt situés autour de l'utilisateur. BrailleNote, également distribué par Humanware, est similaire dans son principe mais permet d'interagir avec l'utilisateur au moyen d'un assistant braille portable plutôt que par le son. Un autre système de guidage relativement populaire est le Kapten (suivi du Kapten Plus). Développé à l'origine pour les cyclistes et randonneurs, il ne comprend pas d'interface visuelle mais différents boutons ainsi qu'un micro pour fournir des commandes vocales, ce qui lui a valu son succès parmi les non-voyants. Enfin, avec le développement des smartphones, qui intègrent maintenant tous des capteurs GPS et des modules de reconnaissance et de synthèse vocale, plusieurs applications mobiles à destination des

déficients visuels ont été développées, comme Loadstone GPS, Mobile Geo, ou Blind Navigator.

Plusieurs évaluations des systèmes commercialisés, tout comme des prototypes développés dans le milieu académique, ont été conduites avec des utilisateurs aveugles ou aux yeux bandés. Le constat de la plupart de ces études est que ces dispositifs ne sont pas suffisamment précis et manquent d'informations pertinentes pour le guidage des non-voyants [Denham et al., 2004; Havik et al., 2010; Strothotte et al., 1996]. La plupart reposent en effet sur des SIG commerciaux, destinés à la navigation automobile. Ils ne contiennent donc pas la plupart des informations nécessaires à un piéton, a fortiori aveugle, telles que la position des trottoirs, des traversées de route, et des autres zones piétonnes. De plus, étant donné que ces systèmes reposent sur la navigation GPS, dont les signaux peuvent être perturbés par les réflexions sur les bâtiments, ou par d'autres facteurs tels que les conditions météo, les performances de positionnement chutent et l'erreur résultante dépasse généralement 20 mètres, pouvant même atteindre jusqu'à 100 mètres.

Pour pallier ce manque de précision, ou l'absence de signaux satellitaires en environnement intérieur, il a par exemple été proposé d'utiliser les signaux WiFi [Hub et al., 2006a], des méthodes de *dead-reckoning,* ou d'améliorer de différentes façon la précision GPS, comme avec le GPS différentiel. Ces points seront abordés dans le deuxième chapitre de ce manuscrit. Une autre démarche consiste à disposer un certain nombre de balises dans l'environnement (infrarouges ou RFID). Celles-ci, si elles constituent un maillage assez dense, peuvent permettre une localisation précise de l'utilisateur [Chumkamon et al., 2008; Mau et al., 2008; Willis and Helal, 2005].

Terminons en mentionnant un dernier type d'aide à l'orientation utilisant également des puces RFID disposées dans l'environnement. Celles-ci permettent d'émettre des messages lorsque un utilisateur se trouve à proximité pour lui indiquer sa position ou la présence de points d'intérêt [Blenkhorn and Evans, 1997; Brabyn et al., 1993; Harris and Whitney, 1995]. Elles ont par exemple été installées dans certaines gares pour aider l'orientation et l'accès aux quais aux non-voyants [Crandall et al., 1999], ou sur des feux de circulation pour indiquer la possibilité de traverser [Zagler et al., 1992]. Evidemment, que ces puces soient utilisées à des fins de localisation ou d'information, le frein majeur à leur généralisation est la nécessité d'équiper l'environnement, ce qui implique des moyens humains et financiers importants pour leur mise en place et leur maintenance.

3.2 Aides basées sur la vision artificielle

Parallèlement aux aides que nous venons de présenter, utilisant des capteurs GPS, inertiels, lasers ou acoustiques, de nombreux systèmes de suppléance utilisant des algorithmes de vision artificielle ont vu le jour depuis l'essor de ce domaine et l'augmentation de la puissance de calcul des dispositifs mobiles. La vision artificielle, ou vision par ordinateur, se définit comme l'ensemble des processus automatisés permettant la compréhension et le traitement d'une scène visuelle. L'utilisation de caméras vidéo, associée à des méthodes de vision par ordinateur, permet donc des champs d'application bien plus larges que les technologies précédemment citées. Il devient par exemple possible de reconnaitre des caractères pour l'analyse automatique de texte, de localiser des objets, d'identifier des visages, de lire des codes-barres, ou de détecter différentes sources potentielle de danger telles que des véhicules en mouvement. Nous proposons ici une revue de ces aides basées sur la vision par ordinateur, en se focalisant sur les deux catégories les plus répandues. La première, à l'instar des dispositifs basés sur les ultrasons ou les lasers, se propose de faciliter la mobilité des non-voyants par la détection d'obstacles, grâce à des caméras embarquées plutôt que par écholocation. La seconde regroupe les dispositifs visant à la reconnaissance et à la localisation d'objets.

3.2.1 Détection d'obstacles

La détection d'obstacles relève de la problématique de la mobilité, c'est-à-dire le fait d'appréhender son environnement immédiat et d'y naviguer de façon sûre. Si la canne blanche ou le chien-guide permettent d'éviter un grand nombre d'obstacles, il reste cependant de nombreuses situations où ils s'avèrent inefficaces, comme par exemple dans le cas d'obstacles élevés tels qu'une fenêtre ouverte ou une branche d'arbre (dans une enquête réalisée auprès de 300 non-voyants [Manduchi and Kurniawan, 2011], plus de 50% des sujets affirmaient être victimes de ce type d'accidents au moins une fois par mois). Certaines approches comme le Télétact ou le Tom Pouce [R. Farcy et al., 2006] proposent d'augmenter les performances de la canne par l'utilisation de faisceaux lasers ou infrarouges. Voir [Roentgen et al., 2008] pour une revue des systèmes électroniques d'aide à la mobilité pour les non-voyants -existant entre 2007 et 2008- qui reposent sur des architectures à base de capteurs laser, infrarouges, ou acoustiques.

L'image, par rapport aux télémètres, nécessite des traitements plus complexes et couteûx, mais constitue une méthode de capture de l'environnement qui offre un éventail de possibilités plus important. Plusieurs projets ont donc tenté de concevoir des systèmes d'aide à la mobilité intégrant des caméras (uniques ou stéréoscopiques).

Un projet portable d'aide aux non-voyants a par exemple été décrit dans [Van Der Heijden and Regtien, 2005]. Il proposait de détecter les obstacles face à l'utilisateur en combinant des cartes de profondeurs calculées au moyen de caméras stéréoscopiques montées sur des lunettes (couplées à une centrale inertielle fournissant les mouvements de la tête), ainsi que et d'un sonar fixé sur la chaussure, balayant l'espace au niveau du sol. Le système étant resté au stade de concept, aucune interface n'a pu être réalisée mais les auteurs suggéraient une interface tactile, sans fournir plus de détails.

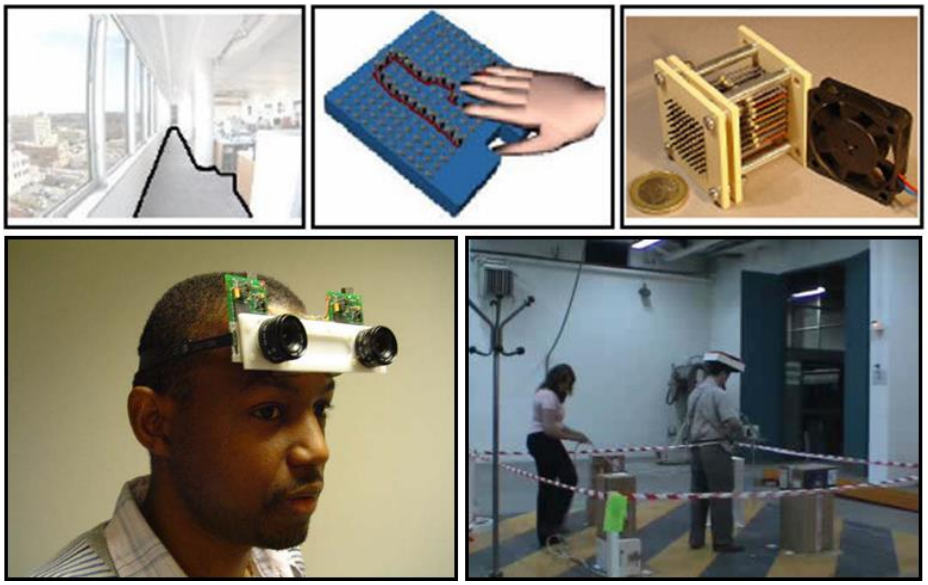


Figure I-39 Projet Intelligent Glasses. La première ligne présente le concept du dispositif, la deuxième le prototype et les expérimentations préliminaires.

Le projet Intelligent Glasses, qui s'est également achevé dans les phases préliminaires, utilise deux caméras stéréoscopiques pour calculer une carte de profondeur de l'environnement [Pissaloux et al., 2008; Velázquez et al., 2004]. Avec l'aide d'un capteur inertiel fournissant l'orientation de la tête, la scène était ensuite projetée orthogonalement sur un dispositif tactile similaire aux tablettes braille pour indiquer la présence d'obstacles relativement à la position du corps. Cette tablette tactile de 8 cm de côté était constituée d'une grille 8 par 8 picots, correspondant à la zone de 4 x 4 m face à l'utilisateur, chacun d'entre eux signalant la présence ou non d'un obstacle dans la sous-région associée. Ce projet, présenté dans la Figure I-39, a été prototypé et a permis quelques expérimentations préliminaires. Si l'architecture du système semble intéressante, l'ergonomie du dispositif, nécessitant les deux mains, présente néanmoins le risque de gêner les comportements de mobilité des non-voyants (qui ne peuvent par exemple pas utiliser une canne blanche ou un chien d'aveugle en complément).

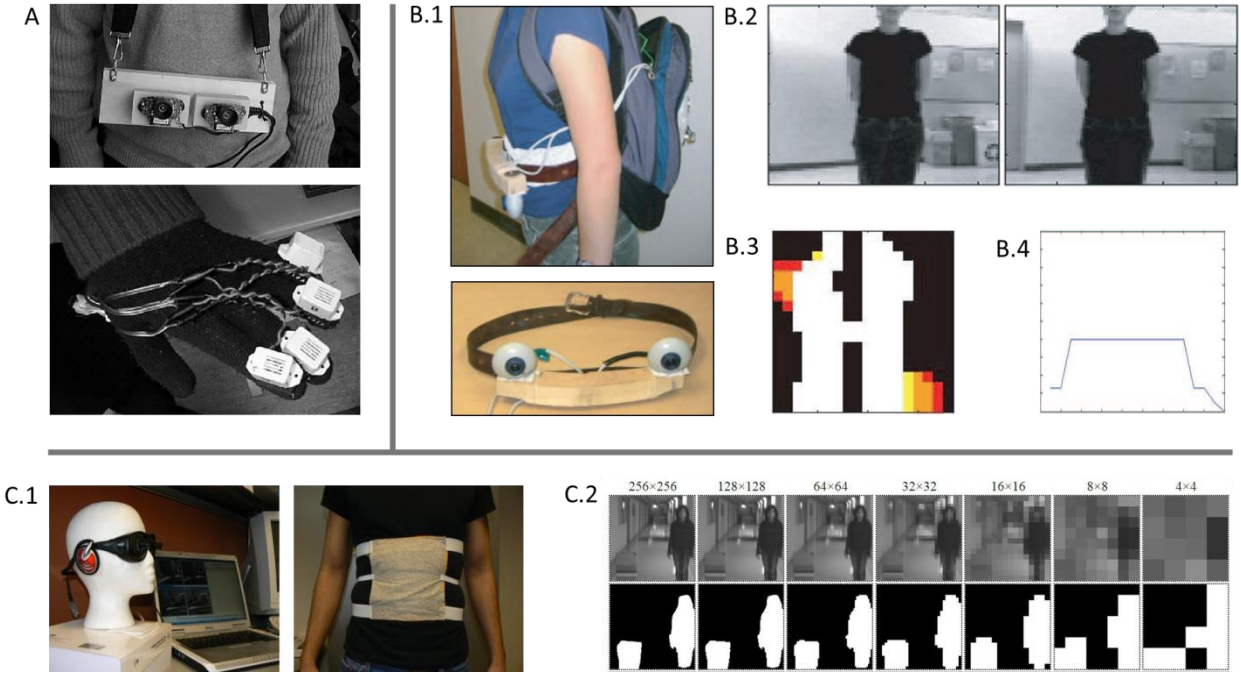


Figure I-40 A) Prototype de l'Université de Guelph [Zelek et al., 1999] ; B) Tactile Vision System [Johnson and Higgins, 2006] (B.1- images du dispositif comprenant une ceinture équipée de deux caméras, un système vibrant sur l'abdomen, ainsi qu'un ordinateur portable dans un sac à dos ; B.2 – images de chacune des caméras ; B.3 - carte de disparité ; B.4 - signaux envoyés à la ceinture tactile) ; C) Dispositif Tyflos comprenant des lunettes stéréoscopiques et une ceinture composée d'une matrice de 4x4 modules vibrants (C.1). Les images et cartes de profondeur sont redimensionnées à cette taille (4 pixels de côté) par une pyramide de sous-échantillonnage illustrée dans C.2.

Plusieurs autres projets utilisent de la même façon une interface tactile et des caméras stéréoscopiques pour indiquer à l'utilisateur la présence d'obstacles proches à partir des cartes de profondeur. Le prototype développé à l'Université de Guelph comprend un dispositif porté autour du cou ainsi que des modules vibrants intégrés à un gant, chacun correspondant à la présence d'un obstacle dans une direction donnée [Zelek et al., 1999]. Le Tactile Vision System (TVS), reposant sur le même principe, décompose la carte de disparité en 14 zones verticales, associées à 14 modules vibrants portés autour de la taille [Johnson and Higgins, 2006], les caméras étant cette fois disposées sur une ceinture.

Une autre méthode de détection d'obstacles a été détaillée dans [Ulrich and Nourbakhsh, 2000]. Si elle s'applique à des robots mobiles et non à l'aide aux non-voyants, elle présente néanmoins l'avantage de la simplicité de l'algorithme et de l'équipement requis : une simple caméra monoscopique. Leur algorithme repose l'hypothèse que l'apparence des obstacles diffère de celle du sol, que celui-ci soit relativement plat, et que les obstacles ne soient pas suspendus en l'air (une partie de ceux-ci doit donc être en contact avec le sol). A partir d'un histogramme de couleurs d'une zone de référence (celle-ci

correspond à une portion de l'image devant le robot ou à des échantillons recueillis dans les trames précédentes, pour lesquels aucun n'obstacle n'a été rencontré), il est alors possible de classifier les pixels de la scène en fonction de leur couleur, comme illustré dans la Figure I-41. Etant donné que l'orientation de la caméra est fixe, plus un élément est loin, plus sa position dans l'image sera haute. Il est donc possible d'estimer la distance des obstacles en fonction de leurs coordonnées verticales. Celle-ci n'étant correcte qu'à leur base, pour chaque colonne, seul le pixel le plus bas sera considéré. Si cette méthode présente de nombreuses limitations (erreurs diverses comme dans le cas des ombres, souvent classées à tort comme obstacles), elle est néanmoins très simple à mettre en place et offre des résultats satisfaisants. L'adaptation pour un piéton serait évidemment délicate, nécessitant une caméra orientée vers le sol, la plus stable possible, et entraînerait probablement des performances moindres, mais pourrait malgré tout apporter des informations utiles en complément d'autres méthodes de détection d'obstacles.

Figure I-41 Classification des obstacles en fonction de la différence de couleur avec une zone de référence [Ulrich and Nourbakhsh, 2000].

Plusieurs algorithmes ont été proposés pour la tâche spécifique de détection des trottoirs et marches d'escaliers. [Se and Brady, 1997] et [Lu and Manduchi, 2005] effectuent pour cela une détection d'arêtes avec un détecteur de Canny et identifient les lignes parallèles caractéristiques grâce à une transformée de Hough. La présence ou non d'un trottoir est finalement prédite autour des lignes parallèles grâce aux cartes de disparité et à l'estimation du plan correspondant au sol. Une autre méthode est employée dans [Pradeep et al., 2008], reposant sur la construction d'une carte 3D de l'environnement à partir de caméras stéréoscopiques et de techniques de SLAM. Les vecteurs normaux en différents points de l'espace sont estimés grâce à l'algorithme RANSAC et au vote de tenseurs (voir

Figure I-42). Un *clustering*[1] de ces normales permet finalement de segmenter les différentes surfaces planaires et ainsi de détecter les trottoirs et les escaliers lorsque plusieurs plans parallèles se trouvent à proximité les uns des autres à des hauteurs cohérentes. Cette méthode, si elle semble offrir de bons résultats, n'a pas encore été testée en conditions réelles, et, par sa complexité, elle nécessite pour l'instant plus d'une dizaine de secondes pour traiter une image de 320 par 240 px, ce qui la rend totalement inadaptée à une utilisation dans un dispositif d'assistance.

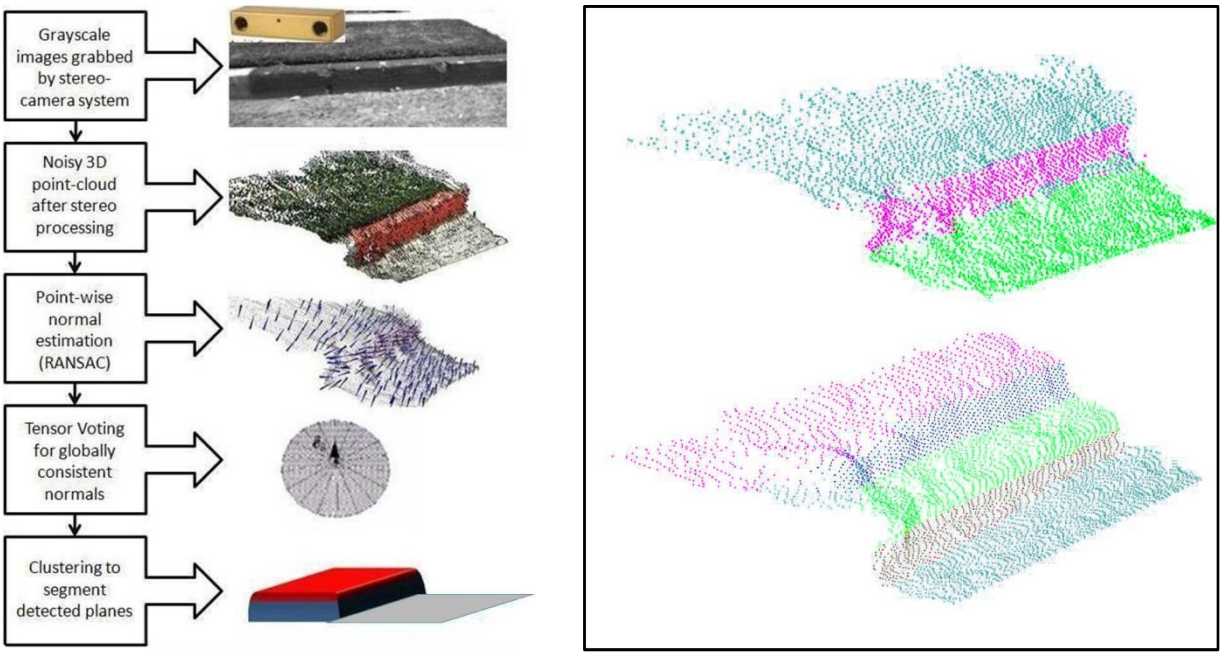


Figure I-42 Détection de trottoirs et d'escaliers [Pradeep et al., 2008]. A gauche principe de l'algorithme, à droite résultat de la segmentation des plans pour un exemple de chaque catégorie (chaque cluster étant représenté par une couleur).

[Martinez et al., 2008] proposent également la construction d'une carte 3D de l'environnement grâce à des caméras stéréoscopiques et des méthodes de SLAM. Le déplacement de l'utilisateur est calculé à partir d'un algorithme d'*egomotion* à 6 degrés de liberté. Cette trajectoire permet de prédire son prochain mouvement pour évaluer les obstacles potentiels. Les caméras étant portées sur l'épaule, et donc soumises à de nombreux mouvements, une méthode de stabilisation a été mise au point pour maintenir la carte locale alignée sur le plan horizontal. Celle-ci se base sur la minimisation d'énergie pour trouver les paramètres de transformation qui conduisent à une entropie minimum dans la distribution 1-d du nuage de points sur l'axe Y. A partir de cette carte stabilisée et du

[1] Le *clustering* est une méthode d'apprentissage non-supervisé, qui consiste à regrouper des données en des ensembles homogènes, appelés cluster.

vecteur de déplacement de l'utilisateur, deux zones virtuelles situées face à lui (à environ 1m50 de distance) sont définies -l'une en position basse, l'autre haute-, et selon leur densité de points respective il est possible de détecter la présence d'obstacles aériens, ou posés au sol (tel un poteau). Cette méthode a pu être implémentée et semble très concluante. Par la parallélisation des algorithmes, l'ensemble des traitements à chaque itération prend environ 364 ms, ce qui permet au système de détecter un obstacle à une fréquence de 2.75 Hz, compatible avec une utilisation en temps-réel.

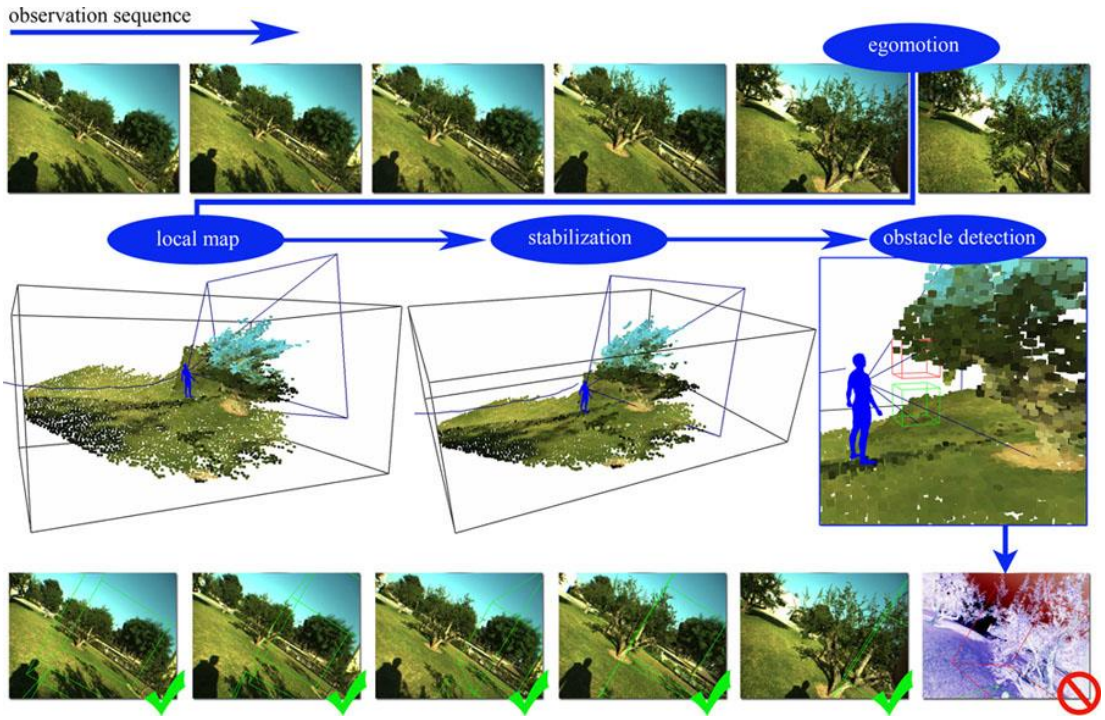


Figure I-43 Architecture de la méthode de détection d'obstacle proposée dans [Martinez et al., 2008]. Celle-ci repose sur la construction d'une carte 3D de l'environnement par odométrie visuelle, puis de sa stabilisation par rapport au plan horizontal, et finalement de la classification des obstacles face à l'utilisateur.

Quelques autres approches peuvent être mentionnées [Adjouadi, 1991; Balakrishnan et al., 2007; Deville et al., 2008; Dunai et al., 2010; Rahman et al., 2004]. Cependant, contrairement aux solutions utilisant des télémètres, très peu des méthodes basées sur la vision ont pu donner des résultats satisfaisants pour l'aide à la mobilité du fait d'interfaces trop complexes, de l'imprécision et des erreurs des cartes de profondeur construites par stéréovision, ou des performances des algorithmes de détections d'obstacles.

3.2.2 Reconnaissance et localisation d'objets

Une autre fonction primordiale pour les aveugles, consiste à reconnaître et localiser des objets d'intérêt dans une scène visuelle. Plusieurs systèmes ont été conçus pour les y aider, permettant par exemple de rechercher un objet spécifique, ou qu'en entrant dans un nouvel environnement, une description des éléments de la scène visuelle soit fournie. On peut regrouper la plupart de ces systèmes en trois catégories : ceux nécessitant un étiquetage, ceux utilisant des modèles 3D, et ceux basés sur les algorithmes 2D classiques de reconnaissance de formes.

Approches à base de tags

La reconnaissance d'objets par ordinateur est une tâche qui, si elle fait l'objet de recherches intensives depuis de nombreuses années, reste cependant très délicate en conditions réelles où le point de vue, l'illumination, ou encore les déformations peuvent rendre l'identification difficile. Pour contourner le problème, tout en utilisant des caméras et des algorithmes de vision artificielle, plusieurs approches utilisent également des étiquettes à apposer sur les objets à reconnaître. Badge3D repose ainsi sur des codes-barres simplifiés. Un filtre de Canny permet de détecter la bande noire les entourant, et donc d'identifier l'objet en analysant leur code-barre. L'utilisateur, équipé d'une caméra sur la tête, peut demander la détection et la localisation d'un objet spécifique grâce à un microphone, afin d'être guidé par des instructions vocales. D'une manière similaire [Gude et al., n.d.] proposent la détection de semacodes[1] sur des objets ou emplacements grâce à des caméras fixées sur des lunettes et sur la canne de l'utilisateur. L'envoi d'informations à l'utilisateur est réalisé au moyen d'une tablette braille. Citons également [Al-Khalifa, 2008], [Nie et al., 2009] et l'application pour Iphone TalkingTag LV[2], qui utilisent tous trois des codes-barres associés à une description audio de l'objet. Notons que ce type d'approches peut également s'appliquer à la navigation. Ainsi, en collant des cercles colorés (voir Figure I-44) sur des endroits d'intérêt (bureaux, toilettes, etc…), Coughlan et Manduchi ont montré qu'il était possible pour un aveugle de s'orienter plus facilement au sein d'un bâtiment grâce à une application sur téléphone portable reconnaissant ces étiquettes [Coughlan et al., 2006; Coughlan and Manduchi, 2007].

[1] Code-barres bidimensionnels : http://semacode.com/about/
[2] http://www.talkingtag.com/lvr/ (application arrêtée depuis Mai 2013)

Figure I-44 Etiquettes de couleurs détectées par une application sur téléphone portable, permettant d'identifier des lieux ou objets d'intérêt [Coughlan and Manduchi, 2007]

Dans ces différentes méthodes, des codes-barres spécifiques ont donc été créés et disposés sur certains objets ou points d'intérêt. Il existe également d'autres solutions, qui visent à lire les codes-barres UPC et MSI[1] existant sur une grande partie des produits de consommation. [Kulyukin and Kutiyanawala, 2010; Kutiyanawala and Kulyukin, 2010] ainsi que [Tekin and Coughlan, 2010] proposent par exemple des applications pour mobile utilisant la caméra du téléphone pour localiser un code-barres dans l'image, analyser celui-ci, et récupérer en ligne des informations sur le produit depuis les bases de données UPC.

Sift, Surf et autres algorithmes de reconnaissance 2D

Il est évident qu'une méthode de détection plus générique, ne nécessitant pas d'apposer des autocollants sur chaque objet à reconnaître, serait une solution beaucoup plus souple et performante. Parmi les nombreuses méthodes de reconnaissance de formes (se reporter notamment au chapitre III), une contrainte importante limitant le choix de l'algorithme est la nécessité d'une vitesse d'exécution (sur dispositif mobile) compatible avec l'utilisation en temps-réel. Si certains projets ont proposé de déporter une partie des calculs sur un serveur distant, ce type d'architecture, malgré l'augmentation de la vitesse des connexions de données mobiles, n'est en général pas compatible avec le traitement d'un flux vidéo [Bigham et al., 2010b]. Ces contraintes excluent donc souvent une grande partie des méthodes de classification reposant sur les SVMs ou les Deep Belief Networks, qui fonctionnent généralement en off-line sur d'importants clusters de calculs. De même, un grand nombre d'approches de segmentation ou d'interprétation complexes de la scène s'avèrent inadaptées. Pour ces raisons, la majorité des systèmes de localisation d'objets pour

[1] Les codes-barres UPC (Universal Product Code) et MSI (aussi appelés *Modified Plessey*) sont des normes de systèmes d'identification largement utilisées sur les produits vendus en magasin.

les non-voyants se sont tournées vers les algorithmes SIFT [Lowe, 2004] et SURF [Bay et al., 2008], réputés pour leur rapidité[1].

Chincha et Tiran ont par exemple développé un système permettant la reconnaissance d'objets du quotidien grâce à une caméra montée sur des lunettes [Chincha and Tian, 2011; Yi et al., 2013]. Les descripteurs SIFT et SURF d'une dizaine d'objets (livre, télécommande, tasse à café, trousseau de clefs,…) ont été pré-calculés à partir d'une base d'apprentissage contenant plusieurs vues de chacun d'eux. Dans la phase de reconnaissance, ces descripteurs sont calculés à partir des images provenant de la caméra et comparés à ceux appris (voir Figure I-45), pour signaler par un son la détection de l'objet recherché. Les résultats des deux méthodes (SIFT et SURF) ont été comparés sur une série d'images test, et si les performances varient beaucoup d'un objet à l'autre (100% d'identifications correctes pour la montre, contre 50% pour la télécommande par exemple), les descripteurs SIFT se sont avérés dans l'ensemble plus précis, mais plus lents que les descripteurs SURF. Les auteurs envisagent par la suite de combiner les deux méthodes dans un framework de Bag-of-Words.

Figure I-45 Détection d'objet par descripteur SURF [Yi et al., 2013]. De gauche à droite : image originale, extraction des points d'intérêt pour le calcul des histogrammes SURF, et matching de ceux-ci dans une nouvelle image.

Le système proposé dans [Cheng et al., 2008] utilise une caméra de 20° d'angle montée sur des lunettes et la détection des objets par les descripteurs SIFT. Deux modes de restitution auditive ont été intégrés, l'un par des sons continus dont la hauteur, le volume et la balance permettent de localiser la cible, l'autre par des instructions vocales indiquant sa

[1] Ces derniers, décris dans le chapitre III, consistent en un appariement de points d'intêret définis par des histogrammes d'orientations

position relativement à l'orientation de la tête de l'utilisateur. Des tests préliminaires du dispositif réalisés avec un sujet non-voyant ont montré qu'il était possible de localiser un objet proche en moins de 10 secondes avec l'interface restituant des sons spatialisés continus et 16 secondes avec l'interface par synthèse de la parole. Néanmoins, si le système fonctionne relativement bien pour des cibles 2D comme les panneaux visibles dans la Figure I-46, pour les objets 3D les performances de reconnaissance sont très aléatoires, y compris dans un environnement contrôlé (à distance fixe et dans des conditions d'éclairage constantes).

Figure I-46 Expérimentation du dispositif proposé dans [Cheng et al., 2008] permettant de localiser des objets grâce à une interface sonore et aux descripteurs SIFT.

Bien qu'il ne s'applique pas à la navigation des non-voyants mais à la surveillance de patients, de personnes âgées ou handicapées à leur domicile, [Xie et al., 2008] proposent également un dispositif de reconnaissance d'objets basé sur les descripteurs SIFT. Dans un souci de réduction des temps de traitements, ils ont inclus un algorithme de détection en cascade basé sur une première passe utilisant des histogrammes de couleurs, très faciles et rapides à calculer, suivi si nécessaire par l'analyse SIFT. Ce type d'architecture, en plusieurs étapes de complexité croissante, semble une alternative intéressante pour compenser les algorithmes de reconnaissance trop lents pour être utilisés dans l'aide aux non-voyants. C'est d'ailleurs ce type de mécanismes que nous proposons dans le troisième chapitre avec le moteur de reconnaissance Spikenet MultiResolution, qui repose sur une première analyse rapide à basse échelle, puis sur une deuxième plus poussée à une résolution supérieure.

Pour terminer, mentionnons deux applications pour iPhone destinées aux non-voyants, LookTel[1] et Vizwiz [Bigham et al., 2010a; Brady et al., 2013], qui reposent respectivement sur les descripteurs SIFT et SURF. S'exécutant sur téléphone portable, les images capturées sont envoyées à un serveur distant effectuant les calculs, ce qui pallie les trop faibles ressources matérielles locales. Dans ces deux systèmes, la création de nouveaux

[1] http://www.looktel.com/recognizer

modèles nécessite le concours d'une personne voyante. Avec LookTel, celle-ci a pour rôle de prendre plusieurs photos caractéristiques de l'objet, et d'y ajouter des annotations (les développeurs mentionnent aussi la possibilité de guidage par un utilisateur distant qui reçoit le flux vidéo et qui indique à l'utilisateur non-voyant comment centrer l'image). A partir de ces images, le serveur ajoute l'objet et ses descripteurs SIFT associés dans une base de données qui peut être partagée par l'ensemble des utilisateurs du système, afin d'offrir un maximum d'objets reconnaissables. Lorsque, lors de l'utilisation, un de ceux-ci est détecté, sa description textuelle est retournée par un moteur de synthèse de la parole. Le système Vizwiz fonctionne de façon légèrement différente. Il permet à l'utilisateur de poser tout type de questions associées à une photo prise par l'appareil (la sélection d'une image correcte, c'est-à-dire nette et correctement cadrée, repose sur un algorithme présenté dans [Zhong et al., 2013], exploitant notamment les accéléromètres de l'iPhone pour s'assurer de la stabilité de la caméra lors de la prise de vue). L'image retenue et la question correspondante sont postées sur des réseaux sociaux tels que Twitter et Facebook, ou sur des services de *crowdsourcing* comme Amazon Mechanical Turk, et les réponses obtenues sont fournies à l'utilisateur par synthèse vocale. Deux autres fonctionnalités sont également proposées par Vizwiz, la première consiste à tenter de reconnaître directement l'objet, sans intervention d'une tierce personne, grâce au service en ligne de Yahoo baptisé IQ Engines, permettant la classification et l'identification automatique d'images. La dernière, LocateIt [Bigham et al., 2010b], combine l'aide à distance et la localisation automatique de l'objet. Pour cela, une photo est envoyée sur un des services précédemment listés afin que quelqu'un segmente dans l'image l'objet recherché. Sa position est ensuite fournie sous forme de sons de fréquences et de hauteurs variables, permettant de guider l'utilisateur. Après l'annotation initiale, celle-ci est déterminée en se basant sur la boussole et le compas du téléphone, ainsi que sur la détection automatique de l'objet au moyen de descripteurs SURF ou d'histogrammes de couleur.

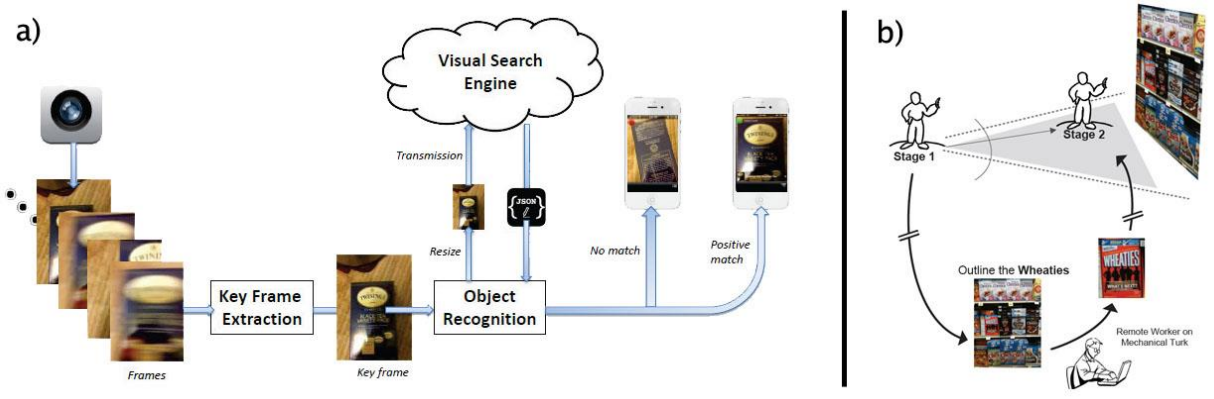


Figure I-47 Application Vizwiz : a) Déportation des traitements de reconnaissance d'objets dans l'application (figure adaptée de [Zhong et al., 2013]) ; b) Module LocateIt permettant la localisation d'un objet à partir d'une segmentation réalisée par des travailleurs en ligne (figure tirée de [Bigham et al., 2010b]).

Modèles 3D

Bien que rares, certaines approches tirent parti de l'information 3D pour la reconnaissance d'objets. Celles-ci supposent généralement un premier scan tridimensionnel de l'objet à rechercher, rendant l'apprentissage non seulement long mais également difficile à mettre en place (l'utilisateur ne possédant ni l'équipement ni le savoir-faire). De plus, ce type de méthode est très coûteux en raison de la complexité algorithmique, et donc souvent inadapté à une utilisation en temps-réel sur un dispositif mobile (aucun système à notre connaissance, n'a adopté une architecture déportée telle que celle que nous venons de le mentionner).

Néanmoins, deux projets ont transposé la reconnaissance 3D à la problématique d'aide aux non-voyants. Le premier, par Kawai et Tomita, combine deux méthodes de stéréovision, l'une basée sur les segmentations, l'autre sur les corrélations, afin d'acquérir une reconstruction 3D de l'environnement [Kawai and Tomita, 2002]. Celui-ci est ensuite confronté aux modèles 3D appris pour la détection puis le tracking des objets trouvés. Des sons correspondants à leur position sont finalement générés au moyen d'HRTF[1] et restitués grâce à un casque stéréo à conduction osseuse. Cependant, comme nous l'avons souligné, ces algorithmes de reconnaissance 3D sont assez couteûx, et ce système est resté à l'étape d'expérimentation offline, les calculs étant trop longs pour être effectués en temps réel.

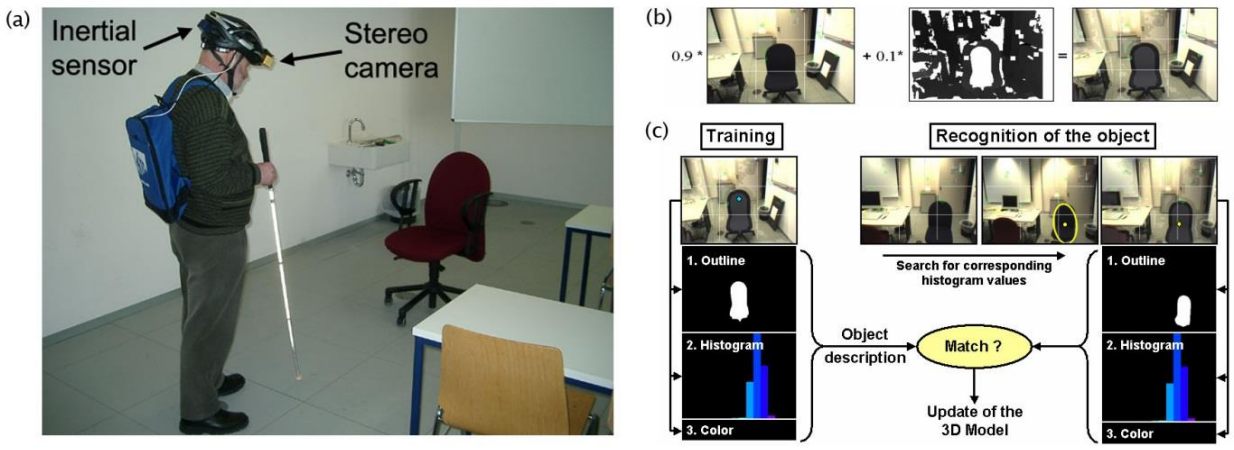


Figure I-48 Suivi d'objets par Hub et al. : (a) dispositif ; (b) fusion de l'image RGB et de la carte de disparités ; (c) algorithme d'apprentissage et de reconnaissance.

Le deuxième projet utilise quant à lui une méthode hybride, et si elle tire effectivement parti d'informations 3D, la reconnaissance des objets n'est pas réellement effectuée en comparant leur modèles géométriques [Hub et al., 2006a, 2006b, 2005]. Le

[1] *Head Relative Transfer Function*, permettant de synthéthiser des sons 3D.

système, prévu pour fonctionner en intérieur, repose sur une modélisation 3D complète du bâtiment, incluant une annotation des objets fixes qui s'y trouvent. La position de l'utilisateur est ensuite déterminée en combinant une localisation par réseaux WiFi et une correction utilisant la carte de profondeurs issue des caméras stéréoscopiques. Par la distance aux murs et par un appariement au modèle 3D de la pièce construit manuellement, il est ainsi possible de déterminer une position et une orientation de l'utilisateur relativement précises. Ensuite, la détection[1] des objets fixes insérés dans le modèle se limite à retourner ceux de la scène virtuelle les plus proches de l'orientation de l'utilisateur. Une autre méthode, principalement 2D cette fois, a été proposée pour les objets mobiles. Elle n'utilise les informations 3D de la carte de disparités que pour segmenter initialement la silhouette de l'objet, au sein de laquelle des histogrammes d'intensités et de couleurs sont calculés en vue de l'apprentissage et de la reconnaissance de la cible, tel qu'illustré dans la Figure I-48.

3.3 Conclusion sur l'approche fonctionnelle

Les dispositifs électroniques d'assistance aux déficients visuels que nous venons de présenter dans cette section ont été créés sur un tout autre principe que les systèmes de substitution sensorielle ou les neuroprothèses détaillés précédemment. Par opposition à l'approche holistique, la démarche consiste ici à répondre à un besoin spécifique des déficients visuels, en mettant à leur disposition un dispositif adapté. De nombreux outils d'aide à la mobilité ont par exemple été développés et commercialisés, consistant pour la plupart à augmenter la canne blanche grâce au principe d'écholocalisation. Ce type de dispositifs semble aujourd'hui la plus mature des aides électroniques au déplacement pour les personnes non-voyantes. Il semblerait que les experts soient même capables de reconnaître certaines formes [Farcy et al., 2003] ou de recueillir des informations sur l'environnement 3D [Hughes, 2001] lors de l'utilisation de ces systèmes.

Pour s'orienter, et se diriger vers une destination, un autre type d'aides a vu le jour il y a une trentaine d'années, baptisées aides électroniques à l'orientation (EOA) ou à la navigation. Ces systèmes fournissent généralement aux utilisateurs des informations sur leur position et les directions à suivre en se basant sur 3 éléments : 1) un module de positionnement (le GPS dans la plupart des cas) ; 2) un système d'information géographique avec une base de données spatiales, et des fonctions de calcul d'itinéraire; et 3) une interface utilisateur qui repose sur une interaction non-visuelle (vocale ou tactile).

[1] Bien qu'il ne s'agisse pas à proprement parler d'une détection.

Enfin, depuis une dizaine d'années, on observe la multiplication des systèmes basés sur la vision artificielle. L'amélioration des algorithmes de vision par ordinateur, et l'explosion des ressources de calcul disponibles sur des systèmes embarqués ont ainsi rendu possible le développement d'aides électroniques aux usages variés. Si ces dernières souffrent encore d'interfaces utilisateur perfectibles, ou de méthodes de traitement des scènes visuelles pas toujours adaptées à une utilisation dans des environnements naturel, les possibilités offertes par la vision artificielle permettent d'espérer l'apparition de systèmes matures dans les années à venir, à même d'aider les déficients visuels dans de nombreuses tâches du quotidien.

4. Synthèse et positionnement

Nous avons vu au cours de cet état de l'art que de nombreuses approches existent pour pallier le handicap visuel : d'une part des aides spécifiques répondant à un besoin identifié, d'autre part des systèmes génériques tels que les systèmes de substitution sensorielle ou les neuroprothèses. Ces derniers ont pour objectif de remplacer ou restaurer la vision dans sa globalité, en tentant de restituer l'ensemble de l'information visuelle au moyen d'une autre modalité sensorielle ou d'une stimulation directe des voies visuelles. L'acuité visuelle et les performances dans des tâches de reconnaissance de formes, de localisation d'objets, ou de mobilité, s'avèrent malheureusement très faibles avec ce type de systèmes. Ils restent donc inadaptés à une utilisation dans la vie quotidienne, indépendamment d'autres problèmes tels que les risques chirurgicaux, le coût des implants visuels, ou la surcharge cognitive des systèmes basés sur des interfaces sonores complexes.

La principale raison de l'inefficacité de ces méthodes est la trop faible résolution de l'interface de sortie. Celle-ci peut être due aux contraintes liées à la modalité sensorielle utilisée par les systèmes de substitution sensorielle (la quantité d'informations exploitables par l'audition ou le toucher étant, comme nous l'avons vu, beaucoup plus limitée que pour la vision), ou au nombre d'électrodes des prothèses visuelles. D'après plusieurs études, il faudrait en effet une résolution bien supérieure à celles des dispositifs existants pour satisfaire aux besoins pour la mobilité [Cha et al., 1992] ou pour l'analyse d'une scène [Pérez Fornos et al., 2008; Zhao et al., 2008]. Comme nous l'avons mentionné précédemment, si la miniaturisation des électrodes pourrait permettre d'augmenter ces résolutions à moyen terme, les gains réels en termes d'acuité visuelle sont beaucoup plus incertains, en raison de facteurs tels que l'incapacité chez les sujets implantés à discriminer les phosphènes résultant de la stimulation de deux points trop proches [Dagnelie, 2008; Dobelle et al., 1974].

Les différents projets de neuroprothèses visuelles, qu'elles soient implantées au niveau de la rétine, du nerf optique ou du cortex visuel, reposent dans leur grande majorité sur une approche appelée *scoreboard.* Celle-ci consiste à reproduire l'image acquise par une caméra à la surface du relais visuel stimulé, en respectant sa configuration spatiale, un pixel correspondant à une électrode. En pratique, cela consiste à redimensionner l'image pour qu'elle corresponde au nombre d'électrodes disponibles, tout en la convertissant en niveaux de gris (sur 2 à 8 valeurs en général, des variations d'intensité plus faibles n'étant pas perceptibles). Différents traitements de l'image peuvent être effectués, comme une détection d'arêtes, un renforcement des contrastes, ou l'application de flou [Hallum et al., 2005] mais ils restent relativement basiques.

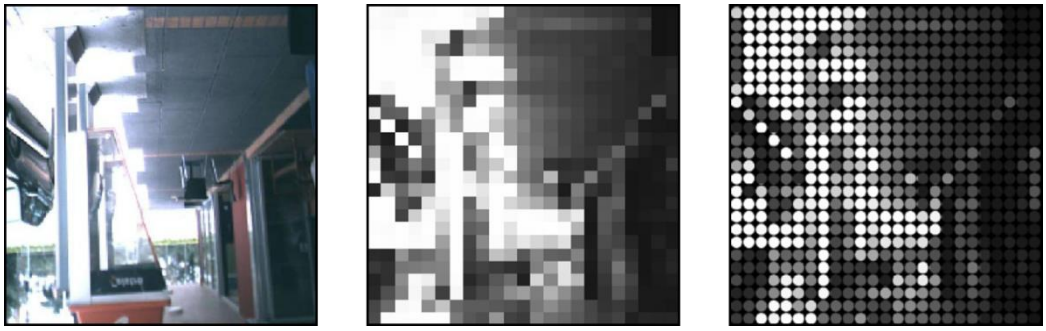


Figure I-49 Simulation d'une neuroprothèse de type *scoreboard*. De gauche à droite : image originale acquise par la caméra ; image convertie en niveau de gris et redimensionnée en fonction de la résolution de l'implant (ici 25 x 25) ; représentation sous forme de phosphènes (images tirées de [Dowling et al., 2004]).

L'approche commune des neuroprothèses et des dispositifs de substitution sensorielle, consistant à retranscrire une image en réduisant sa résolution pour l'adapter à celle de la modalité sensorielle de sortie, nous semble présenter des verrous majeurs au vu du manque de précision des images ainsi converties. Pour concilier la faible résolution des interfaces sonores, tactiles, ou neurales, tout en restituant des informations visuelles pertinentes et exploitables, nous proposons par conséquent dans cette thèse une démarche alternative consistant à prétraiter la scène par un système de vision artificielle.

Afin de valider cette approche, nous présenterons dans le chapitre suivant le développement et l'évaluation d'un système électronique de suppléance pour non-voyants baptisé Navig. Celui-ci s'inscrit dans ce que nous avons appelée l'approche « fonctionnelle », en contraste à l'approche holistique, car elle ne vise non pas à restituer la totalité des informations visuelles, mais à restaurer certaines fonctions spécifiques par une aide adaptée. Tirant parti des possibilités offertes par la vision artificielle, le système Navig permet de réhabiliter une boucle visuomotrice rendant possible la détection, l'identification, et la localisation d'objets d'intérêt. Il repose pour cela sur des algorithmes de vision bio-inspirés exploitant les images de caméras embarquées, ainsi que sur la synthèse de sons virtuels 3D, spatialisés aux coordonnées des cibles détectées. En plus de la localisation d'objets, le système Navig intègre des fonctions de navigation piétonne adaptées aux déficients visuels, permettant ainsi de répondre aux besoins identifiés précédemment en termes de déplacement et d'autonomie. Jusqu'à aujourd'hui, l'utilisation d'aides électroniques à l'orientation a été grandement limitée par l'imprécision du positionnement par satellites (souvent supérieure à 10m, en particulier dans les zones urbaines), qui peut mener à des situations dangereuses (traverser une route en dehors des zones protégées) ou à des erreurs (emprunter la mauvaise rue). Pour compenser la faible précision du GPS, nous proposons

d'utiliser la vision artificielle pour détecter dans l'environnement des cibles visuelles géolocalisées, et raffiner ainsi la position de l'utilisateur. Nous détaillerons cette méthode de positionnement hybride au cours du chapitre suivante, qui combine la détection de points de repères visuels à des données satellites, inertielles, et topographiques, puis nous montrerons les gains apportés dans plusieurs expérimentations réalisées dans un contexte de navigation en milieu urbain.

.

II. Conception d'un système de suppléance basé sur la vision artificielle

Sommaire de section

1. Le projet Navig

Lorsque la vision d'un individu est déficiente, suite à un traumatisme, une maladie, ou de façon congénitale, différentes approches existent pour tenter de restaurer des fonctions visuelles ou de pallier les handicaps générés grâce à des technologies d'assistance. Les systèmes de substitution sensorielle, ainsi que les neuroprothèses, que nous avons détaillés dans le premier chapitre, ont pour objectif de restituer une certaine forme de vision, dans sa globalité, sans prendre en considération les tâches pour lesquelles ils pourraient être employés. Au moyen d'interfaces neurales (implant rétinien par exemple) ou de solutions moins invasives comme des interfaces sonores ou tactiles, la scène visuelle, capturée par une caméra, est transmise au sujet après une conversion relativement sommaire visant à réduire la quantité d'information pour s'adapter à résolution de la modalité de sortie. Comme nous l'avons vu, bien que ces dispositifs permettent de reconnaître quelques formes simples en conditions contrôlées, leur utilisation au quotidien reste très limitée du fait d'une résolution trop faible pour que leurs utilisateurs puissent analyser, interpréter et exploiter les informations visuelles reçues. La complexité et la faible efficacité de ces systèmes expliquent qu'ils n'aient été adoptés que par un nombre très restreint de non-voyants.

En contraste, de nombreuses aides spécifiques ont été développées pour répondre à des besoins précis de la population non-voyante. Dans les domaines de l'accès à l'information, de nombreux systèmes tels que les liseuses d'écran ou les plages tactiles se sont montrés très performants et sont utilisés par de nombreux déficients visuels. A l'heure actuelle, aucun dispositif ne permet en revanche de répondre de façon satisfaisante aux besoins dans les tâches de reconnaissance et de localisation d'objets, ni dans l'aide à la navigation, alors que ces domaines engendrent les incapacités les plus lourdes pour des non-voyants. Etre capable de se déplacer dans un environnement connu ou inconnu, à l'intérieur comme à l'extérieur, soulève en effet de nombreux problèmes liés à l'orientation (savoir où l'on se trouve, se diriger vers une destination voulue,…) ou à la mobilité (éviter des obstacles, maintenir un cap, estimer des distances ou des angles,..).

Le système Navig [Katz et al., 2012a, 2012b; B. F. G. Katz et al., 2010] a pour objectif d'augmenter l'autonomie des personnes déficientes visuelles dans ces deux tâches délicates (la localisation d'objets et la navigation). La conception, le développement et l'évaluation du prototype a rassemblé de nombreux partenaires aux domaines d'expertise variés. Le consortium inclut ainsi deux centres de recherche en Informatique, l'un spécialisé en interactions et technologies d'assistance pour personnes déficientes visuelles (IRIT-ELIPSE), l'autre en perception auditive, cognition spatiale, design sonore, ergonomie et réalité

augmentée (LIMSI), ainsi qu'un laboratoire de Neurosciences de la vision humaine (CerCo), deux PME toulousaines actives en vision artificielle (Spikenet Technology) et en géolocalisation pour piétons (NAVOCAP), un centre d'éducation spécialisée pour déficients visuels (Institut des Jeunes Aveugles - CESDV), et enfin la communauté d'agglomérations du Grand Toulouse.

Ce projet a fait l'objet de trois thèses : celle de Gaëtan Parseihian, au LIMSI (sous la direction de Brian Katz), ayant travaillé sur l'interface sonore, la synthèse des sons spatialisés par HRTF, et les métaphores de guidage [Parseihian, 2012] ; celle de Slim Kammoun, encadré à l'IRIT par Christophe Jouffrais, en charge du contrôleur de dialogue et du système d'information géographique [Kammoun, 2013] ; et enfin la mienne, au sein de l'IRIT, du CerCo, et de la société Spikenet Technology. Afin de clarifier mes contributions, il est important de souligner que bien qu'ayant collaboré avec les différents partenaires sur la plupart des composantes du projet (architecture générale du système, système d'information géographique, guidage, sonification, algorithme de suivi), mon rôle principal concernait les aspects liés à vision. En particulier le système de reconnaissance de formes, la stéréovision, l'apprentissage et la gestion des cibles visuelles à détecter en fonction du contexte d'utilisation, ainsi que le moteur de fusion, qui permet le positionnement hybride à partir des centrales inertielles, du GPS, et de la localisation d'amers visuels. Celui-ci a été développé dans le cadre du stage de Master de Jiri Borovec. J'ai, pour finir, été en charge de l'intégration du système, avec l'aide d'Olivier Gutierrez, alors ingénieur à l'IRIT. Ce travail a consisté à concevoir le prototype, coordonner l'intégration des différents agents logiciels, et développer différents modules annexes comme ceux traitant les données des centrales inertielles ou du GPS, ainsi que des agents de log et de rejeu, permettant de simuler l'envoi des données de chacun des agents du système, enregistrées au cours d'expérimentations, afin de procéder à tests offline.

1.1 Scénarios d'usage

Les fonctionnalités offertes par Navig ainsi que les différents aspects relatifs à l'interface homme-machine seront détaillés au cours de ce chapitre, néanmoins une brève description des scénarios d'usages permettra au lecteur d'appréhender plus aisément la suite du document.

La localisation d'objets et la navigation constituent deux modes de fonctionnement distincts, qui peuvent être initiés et arrêtés à la demande, par une commande vocale ou la pression d'un bouton sur un boitier portable. Pour la localisation d'objet, l'utilisateur doit préciser la cible recherchée parmi la base d'objets ayant été appris (en disant par exemple, « rechercher agrafeuse »). Les descripteurs visuels de l'objet sont alors activés, et si l'objet

est reconnu dans le champ visuel des caméras stéréoscopiques, un son 3D spatialisé sera généré à la position de la détection, puis sera répété aussi longtemps que l'objet continue d'être détecté par le module de vision. L'arrêt de la recherche est commandé par l'utilisateur, ou par le système au-delà d'un temps d'expiration (par défaut si aucune détection n'a eu lieu au cours des 30 dernières secondes). Notons que la cible à localiser peut être un objet au sens classique du terme (un téléphone, une bouteille, un pot de confiture,…), mais aussi tout élément reconnaissable visuellement (un bâtiment, une porte, un visage, etc…). Le second scénario correspond au mode de navigation. Une fois que l'utilisateur a fourni l'adresse de destination, le système va rechercher celle-ci dans les données cartographiques, et, si elle existe, guider le non-voyant à partir de sa position GPS courante. Les métaphores de guidage utilisées reposent également sur des sons 3D, mais aussi sur des instructions vocales. Elles seront abordées dans la suite de section. La fonction de navigation prend fin lorsque l'utilisateur décide d'interrompre le guidage, ou lorsqu'il est arrivé à destination. Ces scénarios d'usage sont illustrés dans un court film de présentation du projet Navig, qu'il est utile de visionner[1] avant de poursuivre la lecture de ce chapitre : *https://www.youtube.com/user/AdrienBrilhault/videos*.

1.2 Architecture générale

Les principaux modules composant le système Navig et leurs interactions sont présentés dans la Figure II-1. Ceux-ci ayant été développés par plusieurs des partenaires listés précédemment, une conception modulaire et cloisonnée a été jugée préférable à un unique logiciel intégré. Elle permet notamment d'utiliser différents environnements de développement pour chacun des composants, de procéder aisément à des tests unitaires, ainsi que de simuler les composants manquants, et elle offre l'immense avantage de permettre d'interchanger très facilement les modules. Pour communiquer, les différents sous-systèmes échangent des messages sur un mode événementiel, par le biais d'un bus logiciel permettant une architecture totalement distribuée. Ce middleware, nommé Ivy, a été développé en collaboration avec le CENA (Centre d'Etude de la Navigation Aérienne), et repose sur l'envoi de messages textuels entre agents [Buisson et al., 2002]. Aucune structure de données complexes ni typées ne peut être envoyée. Cette restriction aux seules chaînes de caractères permet de rendre les données compatibles avec toutes les plateformes et dans tous les langages. De plus, les agents envoient et reçoivent les messages sur une adresse de broadcast. Ainsi, chaque agent peut s'abonner et écouter des messages filtrés par un préfixe de message et invoquer une fonction évènementielle à chaque réception de messages.

[1] De préférence avec des écouteurs stéréo pour une bonne perception des sons binauraux.

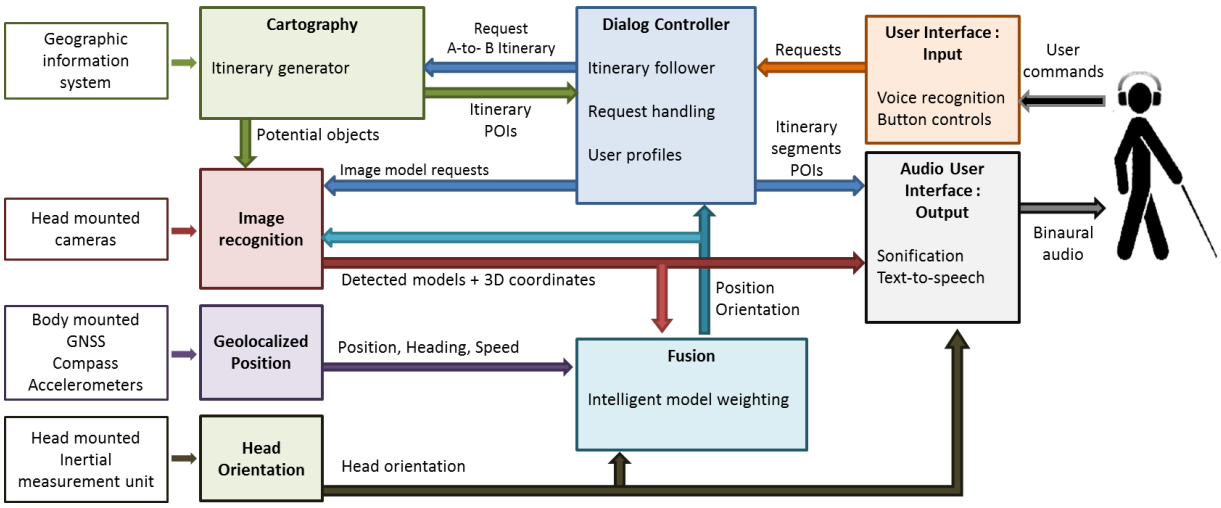


Figure II-1 Architecture générale du système Navig

Ce type d'architecture permet de rajouter, supprimer ou interchanger des agents à la demande, sans altérer le fonctionnement global du système. Des modules réutilisables sont développés de façon indépendante et peuvent être testés, évalués et simulés séparément. On peut ainsi, sans modifier l'organisation du système, substituer un module de synthèse vocale par un autre, changer le système de positionnement, ou encore celui du suivi de mouvements.

Chacun des agents préfixe les messages qu'il envoie sur le bus Ivy par son identifiant, de sorte que les autres modules puissent identifier et filtrer la provenance des messages reçus. Les différents agents du système seront plus amplement détaillés dans la suite de ce chapitre, mais une première vue d'ensemble est présentée ci-après, avec une brève description du rôle de chacun :

- Interface de sortie (IHMS) : effectue la sonification des informations destinées à l'utilisateur
- Interface d'entrée (IHME) : récupère les instructions de l'utilisateur ;
- Récepteur GPS (POS) : détermine la position de l'utilisateur par satellites ;
- Orientation de la tête (CAPT) : fourni les angles yaw, pitch, et roll correspondant à la position du casque ;
- Vision (SNV) : détection des cibles visuelles ;
- Système d'information géographique (SIG) : contient les informations cartographiques et géolocalisées, calcule les itinéraires ;

- Fusion (FUS) : calcule une position consolidée à partir des informations fournies par les agents POS, SNV, et SIG ;
- Contrôleur de dialogue (CD) : coordonne les différents agents du système, et effectue le suivi d'itinéraire et le guidage ;

1.3 Matériel

Le premier prototype du dispositif présenté dans la Figure II-2 consiste en un casque sur lequel est montée une caméra stéréoscopique Bumblebee, commercialisée par Point Grey Research, ainsi qu'une centrale inertielle Mti développée par Xsens Technologies. Celle-ci comprend accéléromètres, gyroscopes et magnétomètres afin de fournir l'orientation de la tête (et par conséquent des caméras) relativement aux 3 axes roulis, tangage et lacet, plus généralement appelés par leur terminologie anglaise, à savoir yaw, pitch, roll. Précisons que le yaw, c'est-à-dire la direction sur le plan horizontal, est donnée relativement au nord magnétique. Dans les prototypes suivants, cette centrale a été remplacée par le modèle Colibri distribué par TriVisio en raison d'une dérive magnétique fréquente observée sur les données du Xsens.

Figure II-2 Prototype Navig comprenant écouteurs, microphone, GPS, caméras stéréoscopique et centrale inertielle montés sur un casque ainsi qu'un ordinateur portable dans le sac à dos.

Le système comprend par ailleurs un microphone pour l'interaction avec le système et la saisie de commandes vocales, ainsi qu'un casque stéréo pour la restitution des sons spatialisés, des instructions de guidage, et tous les autres retours du système destinés à l'utilisateur. Différents types d'écouteurs ont été évalués, et les casques à conduction

osseuse s'avèrent la solution la plus adaptée aux non-voyants [Bruce N. Walker and Lindsay, 2005]. En effet ceux-ci transforment le son en vibrations qui sont directement transmises à la cochlée à travers les os du crâne. Les transducteurs se placent au niveau des tempes (voir Figure II-3) et n'obstruent donc pas le tympan, permettant une interférence minime avec l'environnement sonore ambiant. Le seul problème constaté était un volume maximal trop faible sur certains modèles mais cela est souvent dû à un mauvais positionnement ou à une pression insuffisante.

Figure II-3 Casque à conduction osseuse

Enfin un module GPS fournit la position de l'utilisateur. Il peut s'agir d'un récepteur GPS standard, ou du boitier Angéo, un système de localisation développé par un des partenaires du projet Navig, la société NAVOCAP, qui inclut différents capteurs pour l'amélioration du positionnement. Ces différemment composants sont reliés à un ordinateur portable porté dans un sac à dos (équipé d'un processeur Intel i7 820QM cadencé à 1,73 Ghz et de 4 Go de mémoire).

1.4 Interface utilisateur

L'interaction entre l'utilisateur et le système se divise en deux composantes : l'interface en entrée permettant la saisie de commandes et l'interface de sortie pour restituer les sons et messages. L'ergonomie de celles-ci est cruciale pour que le système soit accepté au quotidien par les utilisateurs, elles ont donc été développées et expérimentées avec le concours de volontaires de l'Institut des Jeunes Aveugles et d'autres professionnels du milieu pour s'assurer qu'elles répondent réellement aux besoin de cette population.

1.4.1 Interface en entrée

L'interface en entrée permet de recueillir les consignes de l'utilisateur. Dans le contexte du système Navig celles-ci peuvent être de différentes natures :

- Saisie d'une adresse de destination ;
- Recherche d'un objet spécifique à localiser ;
- Lancement, interruption ou arrêt de la navigation ou de la détection d'une cible ;
- Gestion du volume sonore ;
- Modification des préférences (palette de sons, stratégie de guidage, de sonification, verbosité, etc.).

Différentes solutions adaptées aux non-voyants sont couramment utilisées dans les systèmes de suppléance, comme les commandes gestuelles (généralement capturées au moyen d'accéléromètres), la saisie clavier (sur téléphone souvent, ou sur des dispositifs spécifiques), ou encore la reconnaissance vocale. Après une phase d'évaluation de ces différentes technologies c'est cette dernière qui a été retenue [Kammoun, 2009]. En effet, les algorithmes de reconnaissance de la parole, bien qu'ils restent en constante évolution et fassent toujours l'objet de recherches, offrent désormais des performances très acceptables pour une utilisation fiable même dans un environnement bruyant. En témoigne leur démocratisation sur de nombreux téléphones portables, tablettes et autres systèmes grand public. Cette modalité d'interaction présente de nombreux avantages de par sa simplicité, son côté naturel ne nécessitant pas ou peu d'apprentissage pour son utilisation, et la facile mise en œuvre grâce aux bibliothèques de développement prêtes à l'emploi et aux faibles contraintes techniques (seul un microphone est requis). Parmi les différentes librairies de reconnaissance vocale, gratuites ou payantes, disponibles sur le marché nous avons opté pour le logiciel *Dragon Naturally Speaking*[1].

Par la suite, une seconde interface, optionnelle, a été ajoutée pour simplifier l'accès aux fonctions basiques comme le volume ou le contrôle du guidage (voir Figure II-4). Celle-ci utilise le boitier Angéo (comprenant le module GPS et différents capteurs pour l'amélioration du positionnement), qui est équipé de plusieurs boutons dont l'activation génère des messages Ivy reçus et traités par le contrôleur de dialogue selon le mapping de fonctions choisi.

Figure II-4 Boitier Angéo

[1] http://www.nuance.com/dragon/index.htm

1.4.2 Interface en sortie

L'interface en sortie, ayant fait l'objet de la thèse de Gaetan Parseihian au laboratoire LIMSI [Parseihian, 2012], utilise la modalité auditive en émettant des messages vocaux (générés par synthèse vocale[1]) et des sons spatialisés. Pour cela le dispositif intègre des écouteurs stéréo à conduction osseuse. Les casques traditionnels couvrent en effet l'oreille et obstruent le tympan de l'utilisateur, limitant l'audition des sons extérieurs. A l'inverse ceux-ci transmettent le son par des vibrations appliquées au niveau des os crâniens, pas des transducteurs situés sur les tempes, permettant au non voyant de continuer à exploiter les sons environnants.

La spatialisation de sons, ou sonification 3D, consiste à générer artificiellement un signal audio stéréo, dit binaural, reproduisant l'impression d'un son émis depuis un point de l'espace particulier. Cette méthode s'appuie sur les perceptions auditives humaines de localisation. Dans le cas de sons naturels, nous sommes en effet capables de déterminer de façon assez précise l'origine de ceux-ci. Cette faculté repose sur l'exploitation par le cerveau de différents phénomènes acoustiques : les différences interaurales, la modification spectrale du son au niveau du pavillon externe de l'oreille, ainsi que les réverbérations dues à l'environnement [Blauert and Allen, 1997].

Le système Navig exploite donc cette capacité de localisation des sons pour fournir un guidage et des informations sur l'environnement spatialisés. Pour cela nous utilisons un moteur de synthèse binaural développé au LIMSI sous l'environnement Max/MSP[2] [B. Katz et al., 2010]. Celui-ci utilise des fonctions de transfert (nommées HRTF pour *Head Related Transfer Function*), qui permettent d'appliquer artificiellement à n'importe quel son les phénomènes acoustiques impliqués dans la perception binaurale et de produire en sortie un flux audio stéréo spatialisé [Begault, 1994; Moller et al., 1995], au moyen d'une convolution par HRIR (*Head Related Impulse Response*). Le dispositif propose deux fonctionnalités principales, la détection d'une cible spécifique initiée à la demande de l'utilisateur, ou le guidage vers une destination. Dans la première tâche la scène sonore se limite à un son unique répété étant spatialisé à l'emplacement de l'objet (cette position, relative à la tête, est calculée par le module de stéréovision). Dans le guidage la sonification est bien plus complexe, incluant des messages vocaux pour indiquer le nom de rues, ou de lieux, des sons 3d pour le guidage le long de l'itinéraire, ainsi que pour des points d'intérêt, etc. Ceux-ci seront abordés dans la partie guidage.

[1] Ou *text-to-speach.*
[2] Max est un logiciel spécialisé dans le traitement du son créé par l'Ircam, permettant la programmation graphique et/ou JavaScript de patchs (http://cycling74.com/products/max/)

1.5 Contrôleur de dialogue

Le contrôleur de dialogue constitue le noyau central du système assurant la coordination des différents agents. Il est en charge de quatre fonctions principales : gérer les interactions avec l'utilisateur, l'itinéraire, le suivi et le guidage. Du point de vue de l'interaction, il réceptionne les messages reçus de l'interface d'entrée, à savoir l'agent IHME (aussi bien vocales que celles du clavier), et procède aux traitements associés. Il lance donc par exemple le chargement des modèles de détection d'un objet que l'utilisateur souhaite localiser par l'envoi d'un message au module vision. Il met également à jour et conserve les préférences de l'utilisateur (pour la planification de l'itinéraire, le choix des points d'intérêts et de repères à présenter ainsi que la verbosité et le type de sons à utiliser en sortie du système), et les transmet aux agents concernés. Lors de la saisie d'une adresse de destination, le contrôleur détermine un itinéraire en fonction de la position de l'utilisateur, de ses préférences de navigation, et du SIG, puis assure le suivi et envoie les instructions de guidage à l'interface de sortie (l'agent IHMS). Ces aspects seront développés dans la partie guidage.

1.6 Système d'information géographique

Un système d'information géographique (noté SIG) est un outil permettant de stocker et traiter des données géographiques. Il doit notamment autoriser des fonctions de capture, de manipulation, d'affichage, de requête et d'analyse d'informations de nature spatiale. Ces deux aspects (stockage et traitement), définis dans [Burrough, 1986; Goodchild, 1991], sont donc organisés en deux composants qui forment l'architecture standard de la plupart des SIG : la base de données géographiques, et le moteur cartographique.

Il existe une grande variété de systèmes d'informations géographiques, pour différents usages allant du guidage automobile (qu'on trouve maintenant dans la plupart des véhicules) à l'épidémiologie, ou encore à l'aménagement du territoire, à l'urbanisme, au marketing, à la géologie, météorologie, etc [Burrough et al., 1998; Clarke et al., 1996; Maliene et al., 2011; Sieber, 2006]. La nature des informations stockées et des primitives d'accès sont donc radicalement différentes selon le domaine d'application (voir par exemple la Figure II-7, extraite de [Burrough et al., 1998]).

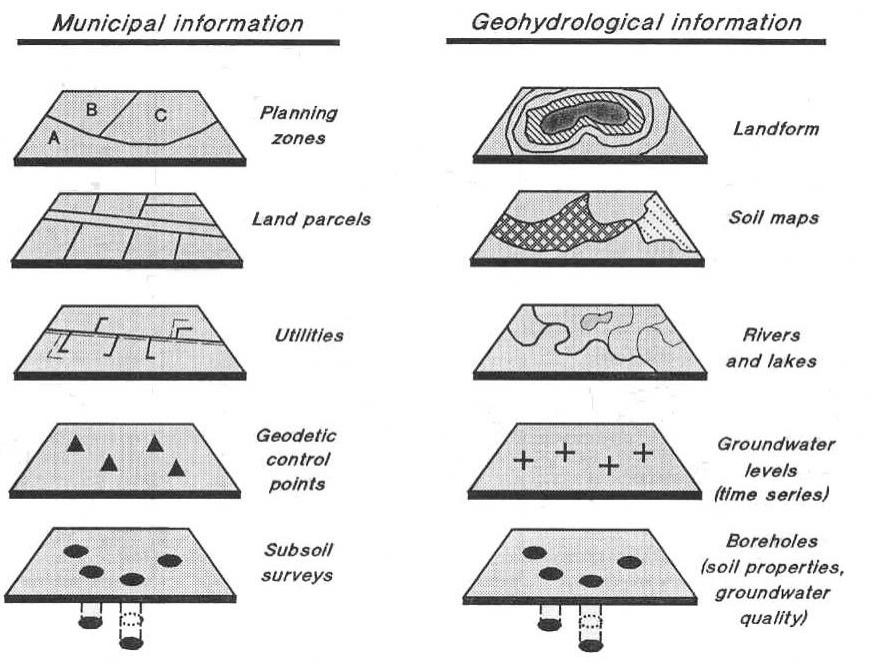


Figure II-5 Exemple de différents types de données géographiques contenues dans des SIG.

Les premiers systèmes d'informations géographiques utilisés pour l'aide à la navigation des non-voyants remontent aux années 80. Le dispositif en question repose sur un positionnement GPS et sur un SIG permettant de calculer un itinéraire, puis d'extraire des points d'intérêt et des descriptions de l'environnement à présenter à l'utilisateur en déplacement [Loomis, 1985; Loomis et al., 1994]. Face au constat que les bases spatiales existantes et les données cartographiques développées à d'autres finalités ne contenaient pas suffisamment d'informations et n'étaient pas assez précises pour le guidage d'un piéton, a fortiori non-voyant, ils ont donc développé un SIG ad-hoc couvrant leur zone d'expérimentations, à savoir le campus de Santa Barbara et ses environs, qu'ils ont rempli à l'aide de la base de données géographiques de l'université et en y incluant les chemins piétons, bâtiments, arbres, et autres obstacles permanents requis [Golledge, 1991].

De nombreuses études se sont intéressées à la question de l'adaptation des SIG pour piétons, déficients visuels ou non [Gaunet and Briffault, 2005; Golledge et al., 2004, 1998; Jacobson and Kitchin, 1997; Zheng et al., 2009]. Il en ressort que les informations indispensables sont les rues, trottoirs, zones piétonnes ainsi que les intersections. Celles-ci ne sont malheureusement pas inclues dans la grande majorité des SIG commerciaux, visant principalement le trafic automobile. Pour guider une voiture, les rues peuvent simplement être modélisées comme des arcs, sans nécessité d'y inclure les passages piétons, trottoirs ou la largeur de la route. Ce type de SIG ne répond donc pas aux besoins de la population non-voyante. Il existe bien quelques projets incluant des SIG à plus ou moins grande échelles

contenant des données géographiques spécifiques aux piétons, comme le système de navigation Navitime au japon, utilisé par près de 2 millions de personnes [Arikawa et al., 2007], ou un service équivalent chinois nommé Navigation Star, mais ceux-ci restent encore très peu répandus.

Dans le cadre du projet Navig, Slim Kammoun a développé un SIG adapté aux non-voyants, incluant ces éléments de base mentionnés précédemment ainsi que d'autres types de données jugées utiles pour la navigation des aveugles [Kammoun et al., 2012]. Ceux-ci sont les résultats des interviews et des méthodes de conception participative mises en place avec l'Institut des Jeunes Aveugles de Toulouse, ainsi que des conclusions de différentes études analysant leurs besoins [Gaunet and Briffault, 2005; Golledge et al., 2004]. Une fois collectées et stockées dans le SIG elles seront prises en compte dans la planification d'itinéraire et pourront ensuite être fournies à l'utilisateur pendant la préparation de l'itinéraire ainsi que pendant la navigation. Les différentes catégories d'éléments intégrés dans ce SIG sont présentées ci-dessous :

- Zones piétonnes : Elles comprennent l'ensemble des aires utilisables par un piéton, à savoir les trottoirs, passage piétons, parcs, et autres aires où il est possible de se déplacer.
- Points de repère (PR) : Lieux ou objets pouvant être détectés par une autre modalité sensorielle que la vision. Ces points doivent permettre à l'utilisateur de confirmer sa propre position dans le trajet. Cet aspect est important pour que l'utilisateur puisse se fier au système. En effet en cas d'imprécision du GPS, un voyant peut vérifier qu'il se trouve bien sur le trajet indiqué en regardant les noms de rue par exemple, un déficient visuel en revanche n'aura pas cette possibilité de confirmation et c'est donc le rôle de ces points de repère. De plus, si l'utilisateur est amené à refaire cet itinéraire, il pourra de nouveau se repérer grâce à ceux si, sans nécessairement l'aide du dispositif.
- Points d'intérêts (POI) : Lieux présentant un intérêt potentiel pour l'utilisateur. Ils peuvent être utilisés comme destination finale ou juste être signalés à l'utilisateur pendant son déplacement afin de lui permettre une meilleure compréhension de l'environnement (e.g., bâtiments publics, magasins, ...). Ces points sont accompagnés d'une catégorie ainsi que d'une description pouvant être fournies à l'utilisateur. A titre indicatif, la première version de notre système intègre 7 catégories différentes : restauration, hôtellerie, commerces divers, commerces alimentaires, services, sport/loisir, et tourisme.
- Points difficiles (PDF) : Ils correspondent aux traversées, aux intersections ou carrefours ainsi qu'à tous les passages du trajet qui peuvent être considérés comme

problématiques pour les non-voyants et dont l'utilisateur doit être informé de façon spécifique. Ils ont pour fonction d'alerter l'utilisateur.

- Points favoris (PF) : Il s'agit de points d'intérêt définis par l'utilisateur comme son lieu de résidence, un commerce, l'adresse d'un ami,... Ils peuvent être ajoutés dans le SIG au cours de ses trajets ou hors-navigation au moyen d'une commande vocale spécifique.
- Points de vision (PV) : Ces points, que j'ai mis en place au cours de cette thèse, constituent des amers visuels utilisés par le module de vision. Il s'agit de cibles pouvant être détectées au cours du trajet par les caméras afin de raffiner l'estimation de la position de l'utilisateur. Leur rôle et leur nature seront développés dans la section suivante.

Un point de l'environnement peut évidemment appartenir à différentes catégories. Un exemple d'un trajet simple et des différents points pouvant être rencontrés est donné pour illustrer cette classification dans la Figure II-7 (tirée de [Kammoun, 2013]). Plusieurs solutions ont été envisagées pour constituer automatiquement le SIG du système. Il existe en effet de nombreuses bases de données libres, ou propriétaires :

- Bases commerciales : Parmi les bases commerciales les deux plus populaires sont Navteq et TéléAtlas, rachetées en 2008 par Nokia et TomTom, qui comportent une représentation détaillée du réseau routier.
- Google Maps : Google propose également des services de cartographie permettant la visualisation de cartes, et le calcul d'itinéraires (pour voitures, piétons et vélos), néanmoins leur base ne comprend pas les trottoirs, passages piétons, et autres éléments requis pour un guidage convenable.
- Open Street Map : Le projet OpenStreetMap propose une base libre, modifiable de façon communautaire à la manière des wikis. Avec un grand nombre d'utilisateurs, celle-ci s'est rapidement enrichie de nombreuses annotations de bâtiments, parkings, restaurants, routes, et comporte même les trottoirs sur certaines régions.
- Autres bases : en plus de ces bases aisément accessibles (qu'elles soient ou non gratuites), beaucoup de collectivités possèdent leur propre SIG. L'agglomération toulousaine dispose par un exemple d'une base de données très détaillée nommée Toulouse Métropole. Elle inclut la grande majorité des éléments de la voirie (trottoirs, bouches d'égout, plots, barrières, poteaux, et autre mobilier urbain), et nous a été mise à disposition par le Grand Toulouse, membre du consortium Navig.

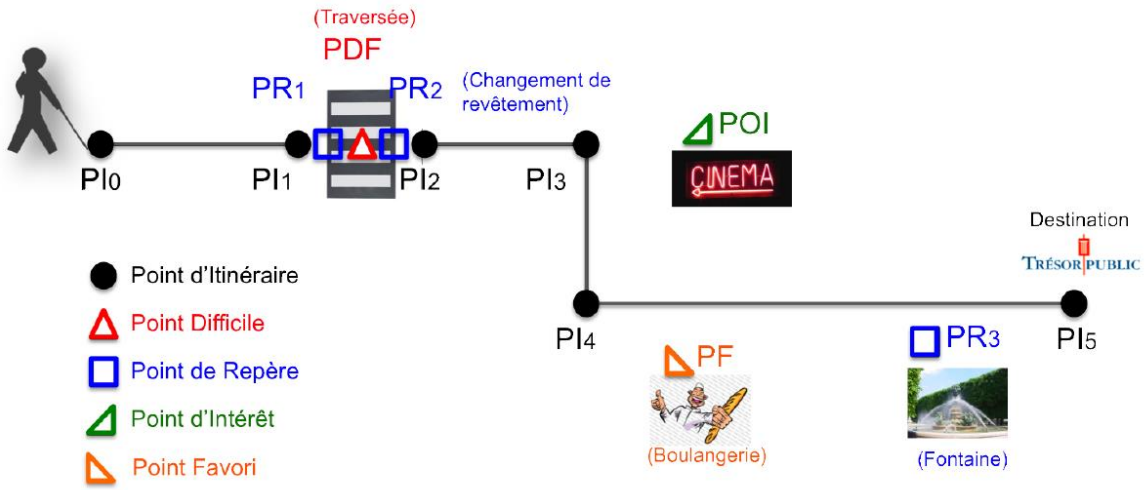


Figure II-7 Différente catégories de points extraits du SIG au cours d'un trajet.

Aucune de ces différentes bases de données géographiques ne contenant l'ensemble des informations nécessaires pour un système d'aide à la navigation des non-voyants, nous avons collecté celles-ci en nous limitant aux lieux sur lesquels ont été effectuées les expérimentations du dispositif, c'est-à-dire le campus, et le quartier des Carmes dans le centre-ville. Nous sommes partis des informations extraites d'OpenStreetMap, du SIG de l'agglomération toulousaine et de celui de l'Université Paul Sabatier, auxquelles nous avons ajouté manuellement les éléments manquants grâce à l'éditeur spécialisé Quantum GIS. Dans sa thèse, Slim Kammoun propose une description détaillée des spécifications de cette base de données [Kammoun, 2013], dont un modèle conceptuel est illustré dans la Figure II-6.

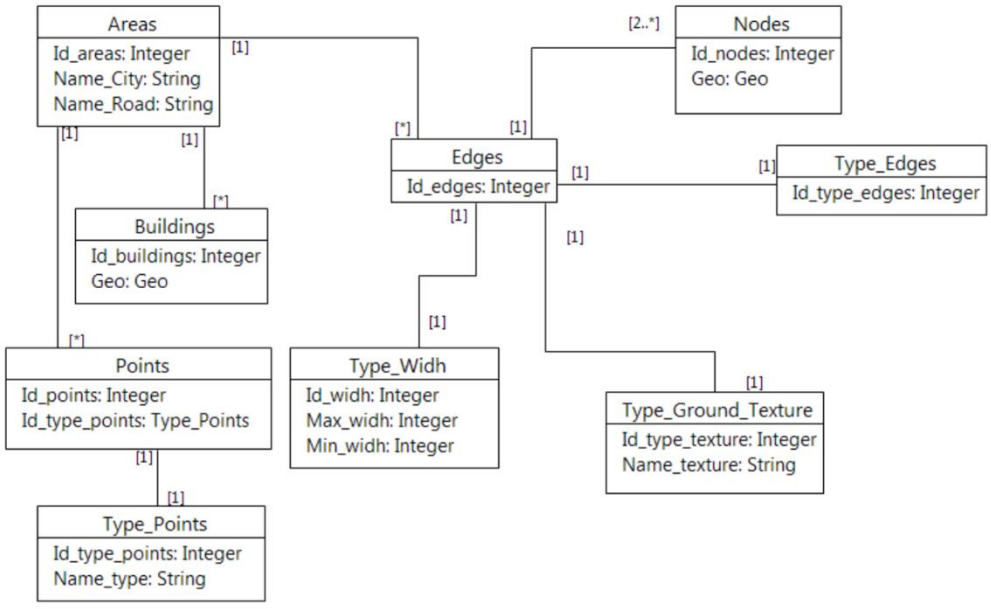


Figure II-6 Modèle conceptuel de données du SIG Navig [Kammoun, 2013].

Du point de vue de l'extraction de ces données, un moteur cartographique a été développé proposant la plupart des requêtes spatiales généralement incluses dans les SIG. Ces différentes primitives d'accès sont fournies ci-dessous :

- Où suis-je : Cette fonction assure la conversion d'une position géographique sous la forme d'une adresse (numéro, rue et lieu).
- Calcul d'itinéraire : Détermine un itinéraire (sous la forme d'un ensemble de segments appelé linéaire) entre deux positions.
- Liste des points dans un cercle : Permet de retrouver tous les points marqués dans la base de données comme POI, PR ou PV dans un cercle dont le centre et le rayon sont fournis en paramètre de la requête.
- Liste des polygones dans un cercle : Identique à la précédente fonction mais retourne l'ensemble des polygones dans un cercle.
- Liste des lieux dans un disque : Retourne de façon similaire l'ensemble des lieux (c'est-à-dire les noms de rues, places, etc.)
- Liste des points dans un linéaire : Renvoie la liste des points autour d'un linéaire selon une distance définie.
- Liste des bâtiments dans un polygone : Recherche l'ensemble des polygones existants dans un polygone donné.

En plus de ces fonctions d'extraction, trois fonctions ont été proposées pour ajouter, modifier ou supprimer un point. Bien que cela dépasse le cadre de ces recherches, soulignons les perspectives de mise à jour participative de cette base. Les points favoris inclus dans notre système offrent par exemple cette possibilité, en autorisant l'utilisateur à ajouter à la demande de nouveaux POI dans la base. D'autres méthodes mentionnées dans [Völkel and Weber, 2007] et [Yan et al., 2009] pourraient aussi être appliquées. S'appuyant sur la théorie du contrôle par feedback (*Feedback Control Theory*) des systèmes dynamiques, Yan et al. proposent par exemple un modèle où le piéton n'est pas seulement récepteur d'information, mais est aussi considéré comme une source de données, de façon explicite ou non. Il peut par exemple ajouter de nouveaux points d'intérêt qui seront partagés avec les autres utilisateurs, assigner des notes de difficultés à différents segments d'un trajet emprunté, et ses déplacements ainsi que leur vitesse peuvent aussi être exploitées pour la mise à jour automatique des aires piétonnes, ou des calculs d'itinéraires si certaines rues, de par leur fréquentation ou leurs obstacles s'avèrent plus longues à parcourir.

1.7 Calcul et suivi d'itinéraire

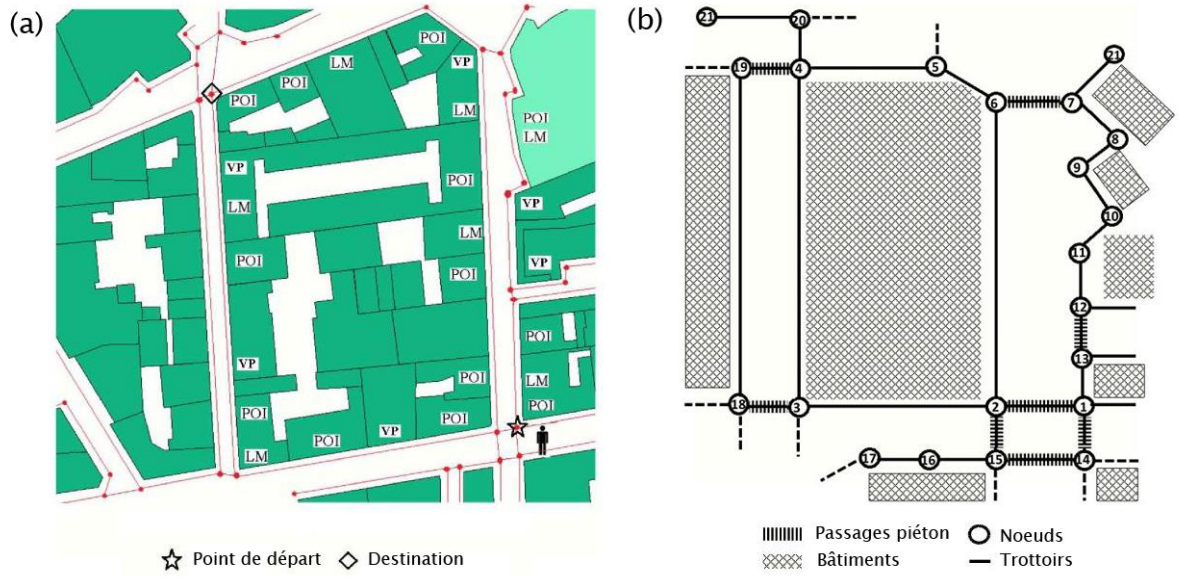


Figure II-8 Représentation d'une zone du centre-ville sous forme de graphe pour le calcul d'itinéraire (avec annotation des points autour du trajet à gauche, et sous forme simplifiée à droite).

Le calcul d'itinéraire consiste à déterminer un trajet entre deux points de la carte, sous forme de segments connexes. Les chemins utilisables par un piéton étant modélisés dans le SIG sous forme de graphe, il s'agit donc de calculer le plus court chemin entre deux nœuds de celui-ci. Pour cela la plupart des méthodes de calcul d'itinéraires reposent sur l'algorithme de Djikstra [Dijkstra, 1959]. Il existe près d'une vingtaine d'algorithmes différents permettant de résoudre ce type de problème, évalués dans différentes études, aussi bien sur des graphes simulés que sur de réelles données cartographiques [Cherkassky et al., 1996; Zhan, 1997; Zhan and Noon, 1998], il en ressort que l'algorithme de *Dijkstra's Approximate Buckets* offre de très bon résultats et c'est donc la méthode qui a été implémentée par Slim Kammoun dans le SIG Navig [Kammoun et al., 2010].

Un poids (ou score) est associé à chacun des arcs du réseau. Si l'on souhaitait calculer le chemin le plus court en termes de distance à parcourir il suffirait de choisir comme score la longueur de l'arc. Cependant les entretiens réalisés avec des membres de l'Institut des Jeunes Aveugles nous ont montré que les non-voyants peuvent souvent préférer un itinéraire plus long s'il comporte moins d'obstacles ou de difficultés. Le choix du trajet varie donc en fonction de l'utilisateur (suivant son expérience en mobilité) et du type de déplacement (milieu connu ou inconnu, balade ou rendez-vous, etc.).

Nous avons, en plus de la distance, intégré différents critères relatifs aux points difficiles et aux autres catégories de points du SIG présentés précédemment. Des pénalités sont associées aux points difficiles, aux traversées de rue, à la largeur des trottoirs et aux intersections, en fonction de leur nombre d'embranchements, voir [Haque et al., 2007]. A l'inverse, des « bonus » (ou profits) sont calculés en fonction de la présence de POI, de PR et de PV proches de l'itinéraire. Ceux-ci sont en effet utiles à l'utilisateur pour se créer une représentation de l'environnement, pour confirmer sa position, ou permettre au système de corriger les coordonnées GPS dans le cas des Points Visuels.

Le choix des coefficients de pondération doit être personnalisable selon les utilisateurs ou bien l'environnement dans lequel la tâche de navigation se déroulera. Par exemple, dans un centre-ville dense où le GPS n'est pas assez fiable, un poids plus important sera accordé aux cibles visuelles pour compenser la précision de positionnement. De plus, si l'utilisateur est en mode découverte par exemple, et aimerait avoir plus d'informations sur les POI, on augmentera le score associé à ce type de points. Pour un trajet utilitaire on préférera en revanche privilégier la distance la plus courte. Une présentation de ces différents choix de pondération et de normalisation est proposée dans [Kammoun et al., 2010], dont est extraite la Figure II-8.

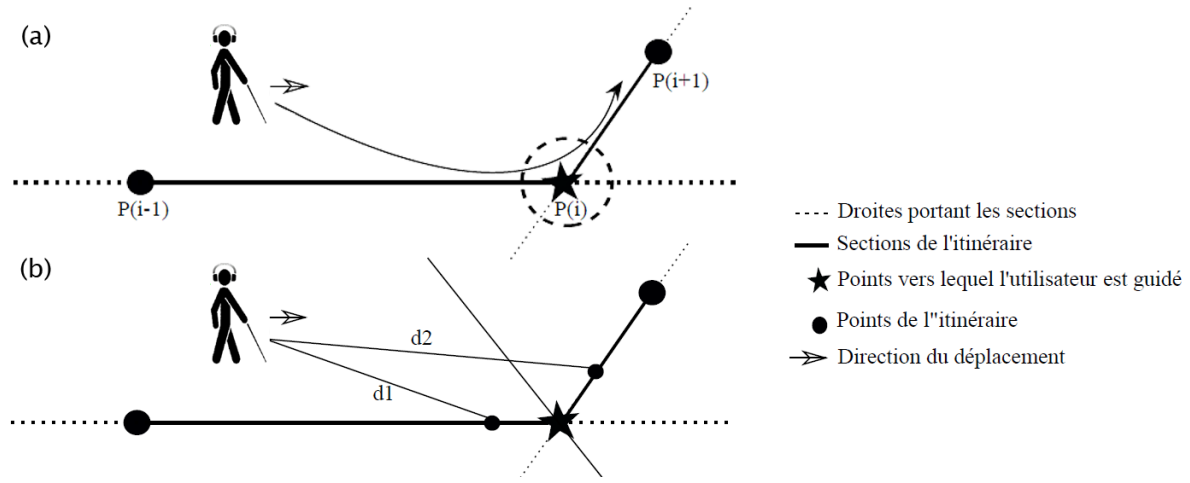


Figure II-9 Algorithmes de suivi d'itinéraire : (a) méthode du rayon autour des points de passage ; (b) méthode de la bissectrice.

Une fois le calcul d'itinéraire effectué par le SIG, celui-ci est retourné au contrôleur de dialogue sous la forme d'une succession de points d'itinéraire, de points difficiles, ainsi que des points de repère, d'intérêt, visuels et favoris présents à proximité du parcours. Le guidage peut alors commencer, et le contrôleur de dialogue doit tout au long de celui-ci déterminer l'avancement du trajet, c'est-à-dire positionner l'utilisateur sur l'itinéraire afin de savoir vers quel prochain point de passage l'orienter, quelles informations fournir, ou encore procéder à un nouveau calcul d'itinéraire s'il s'en éloigne trop. On appelle généralement

cette tâche le suivi d'itinéraire. Différentes stratégies existent pour traiter ce problème, beaucoup de systèmes de navigation pour non-voyants reposent sur une simple distance aux points de passage pour valider ces derniers [Walker and Lindsay, 2006]. Cette approche pose néanmoins des problèmes en cas d'imprécision du positionnement. Si celle-ci est supérieure au rayon choisi (généralement entre un et deux mètres), l'utilisateur court le risque de ne pouvoir le valider et donc de continuer son trajet [Ivanov, 2011]. L'agrandissement du cercle autour de chaque point n'est pas non plus une solution satisfaisante, car les instructions de guidage seraient alors données bien trop tôt par rapport au réel changement de direction. Nous avons donc proposé un nouvel algorithme reposant sur la distance relative à deux points temporaires créés à distance égale le long de l'itinéraire avant et après le point de passage, tel qu'illustré dans la Figure II-9. Le point est ainsi validé une fois franchie la bissectrice de l'angle formé par ces deux points et le point de l'itinéraire. La souplesse de cette méthode permet un guidage optimal quelle que soit la précision du GPS [Kammoun, 2013].

1.8 Guidage

Une fois le trajet initié le rôle principal d'un système d'aide à la navigation est de guider l'utilisateur de manière sûre et fiable. En fonction de leur préférence, de leur expérience en mobilité, et de leur familiarisation avec le système, les utilisateurs du dispositif ont la possibilité de choisir différents niveaux de détail quant aux informations qui leur seront présentées. Les différents entretiens et *brainstorming* effectués avec un panel de non-voyants ont permis d'identifier deux types de navigation à prendre en compte. Le mode standard se contente de guider l'utilisateur vers sa destination, le plus rapidement et efficacement possible. Par conséquent seules les informations strictement nécessaires à la réalisation de cette tâche lui sont fournies, à savoir les points de l'itinéraire (PI), les points difficiles (PDF), et les points de repère (PR). A l'inverse, dans le mode exploration, l'utilisateur cherche à découvrir un quartier ou un trajet spécifique. Il convient donc de restituer un maximum de description de l'environnement, en signalant par exemple les commerces, bâtiments, arrêts de bus, etc. En plus des points du mode standard sont donc ajoutés les points d'intérêt (POI), que l'utilisateur peut personnaliser en fonction des informations désirées grâce aux sous catégories mentionnées dans la partie SIG (restauration, commerces, tourisme,…).

Les points de l'itinéraire et la destination finale sont toujours représentés par un son 3D positionné à l'emplacement du prochain point de passage, dont la fréquence de répétition est paramétrable (en secondes ou en mètres). Les autres points en revanche (POI, PDF et PR), peuvent être indiqués par un vocabulaire sonore sémantique et/ou synthèse

vocale selon le niveau de verbosité choisi [Parseihian, 2012]. Ces informations seront dans tous les cas restituées par sonification spatialisée, de sorte que la description de chaque objet semble provenir de sa position réelle. Pour cela les descriptions textuelles ont pu être générées en couplant la librairie de *text-to-speach* Acapela au moteur de sonification 3D.

Le vocabulaire sonore utilisé pour les différents types de points à signaler à l'utilisateur est basé sur une extension du concept d'*earcons*, des motifs de notes, qui, structurés selon une certaine grammaire, permettent de transmettre de l'information de façon non-verbale [Blattner et al., 1989]. Fonctionnant sur le même principe, les *morphocons* sont des motifs de paramètres acoustiques (tels que la fréquence, le tempo ou la dynamique) permettant la construction d'un langage sonore hiérarchique basé sur la variation temporelle de ceux-ci [Parseihian and Katz, 2012]. Il est donc ensuite possible, une fois défini un ensemble de morphocons, de les appliquer à plusieurs ensembles de sons pour créer différentes palettes sonores que l'utilisateur pourra sélectionner selon ses préférences sans réapprentissage (les formes sonores pouvant être reconnues indépendamment du contexte [Frey et al., 2009; Minard et al., 2010]).

Dans le cadre du projet Navig un vocabulaire de *morphocons* a donc été développé pour permettre aux utilisateurs de rapidement identifier et différencier chaque classe d'objets [Parseihian, 2012]. Il répond aux attentes exprimées par le panel de non-voyants interrogés, préférant généralement des sons brefs (pour éviter le masquage des sons réels, la surcharge cognitive et la fatigue auditive), plaisants (proscrivant donc les sons type bruits blancs ou tons purs), et facilement discriminables des bruits de l'environnement. Il tient également compte des contraintes techniques du dispositif en réduisant au possible les variations dynamiques difficilement perceptibles avec des casques osseux, et en générant des sons à spectre large et aux attaques franches, connus pour améliorer la perception des indices de spatialisation.

Au total le vocabulaire créé comprend cinq catégories (PI, PDF, PR, POI et PF) illustrées dans la Figure II-10, et 13 sous-catégories (4 PR, 7 POI et 2 PF). Les PI sont décrits par un son bref, les PDF par deux successifs, les PR par trois, dont le motif rythmique permet de différencier les 4 sous-catégories. Enfin les POI et PR sont symbolisés par deux groupes de sons consécutifs. Le premier étant commun à tous les éléments (un son dont la fréquence augmente pour les POI, et décroit pour les PF), et le second pouvant être constitué d'un ou plusieurs sons courts pour distinguer les sous catégories (par exemple un son grave, un son aigu, deux notes ascendantes ou descendante, 3 notes ascendantes puis descendantes, etc.). La durée des *morphocons* utilisés varie de 200 ms à 1,5 s. A partir de ce vocabulaire trois différentes palettes de sons ont été développées et évaluées par 30 sujets (31 voyants et 29 non-voyants) :

- La palette naturelle constituée principalement de cris d'oiseaux ;
- La palette instrumentale comprenant des sons d'instruments à cordes ;
- La palette électronique réalisée à partir de sons de synthèse.

L'analyse des résultats des tâches de classification a montré de très bons taux de discrimination entre les catégories principales (78 ± 22 %), sans différences significatives entre voyants et non-voyants, et de bonnes performances, bien qu'un peu en deçà, pour les sous-catégories (63 ± 23 % pour les POI, 58 ± 29 % pour les PR et 87 ± 19 % pour les PF). Cette étude a permis de pointer des variations rythmiques trop proches pour les sous-catégories de PR expliquant leur taux de reconnaissance plus faible, ainsi que des problèmes spécifiques à certains sons dans les différentes palettes [Katz et al., 2012b; Parseihian, 2012; Parseihian and Katz, 2012]. Sur la base de ces constatations, trois nouvelles palettes sonores sont en cours de développement pour les versions futures du système Navig.

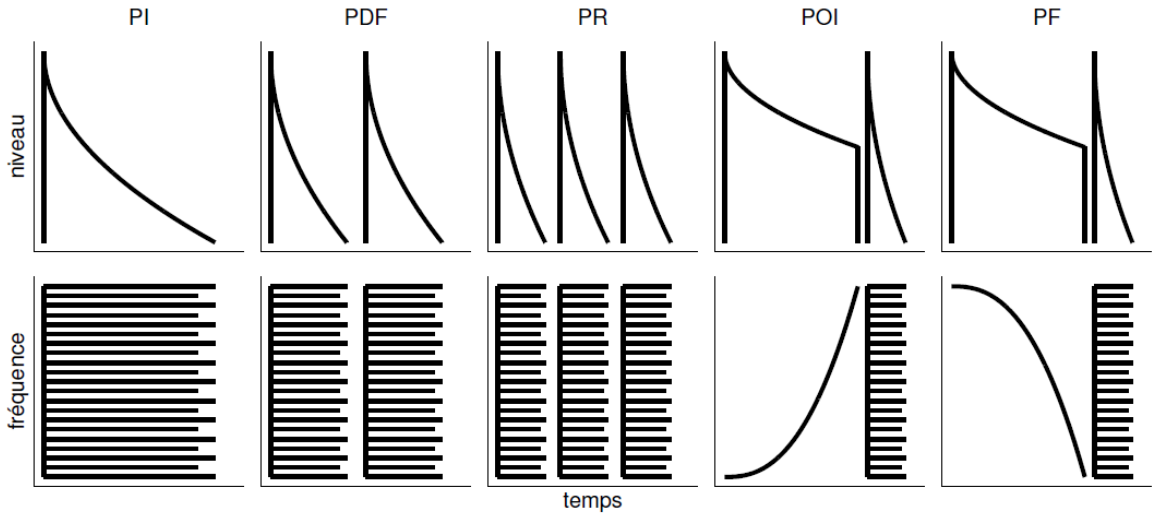


Figure II-10 Profils des morphocons utilisés pour chaque type d'objet sonore du système Navig.

2. La vision dans Navig

La localisation d'objets constitue évidemment la fonction la plus attendue d'un système d'aide aux non-voyants basé sur la vision par ordinateur. Il s'agit donc de détecter et localiser des objets d'intérêt dans les flux vidéo acquis par des caméras embarquées, dont la position pourra être ou non restituée à l'utilisateur selon le contexte. On distingue en effet deux modes distincts de fonctionnement du système Navig. Le premier permet à l'utilisateur de demander à tout moment la recherche d'un objet particulier ayant été préalablement appris par le système, tel qu'une boîte aux lettres, un distributeur automatique, une bouteille ou encore un téléphone, et ce afin de s'en saisir, de se diriger vers celui-ci, ou simplement d'en connaître la position. Le deuxième mode de fonctionnement correspond à une tâche de navigation. Ici le non-voyant cherche à rejoindre une destination, et sera guidé tout au long du trajet à partir des informations cartographiques et du GPS. Néanmoins, même en navigation, différentes cibles visuelles géolocalisées sont recherchées par le module de vision de façon transparente pour l'utilisateur, dans le but de raffiner la précision du positionnement en cas d'erreur du GPS. Chacun de ces deux aspects est développé dans cette section.

2.1 Traitements visuels

Quel que soit le mode de fonctionnement du système, les traitements visuels, matériels comme logiciels restent absolument identiques. Ceux-ci reposent sur une ou plusieurs caméras permettant de capturer l'environnement autour de l'utilisateur, et sur le moteur de reconnaissance de formes Spikenet présenté dans la suite de ce chapitre.

2.1.1 Localisation 3D

Si l'algorithme de reconnaissance est capable de localiser une cible visuelle particulière, cette détection reste dans l'espace 2D de l'image, et ne donnera par conséquent que la direction de l'objet reconnu dans le référentiel de la caméra (donc de l'utilisateur, étant donné que les caméras sont montées sur un casque), et non une position 3D. Or dans le contexte d'un dispositif d'aide aux déficients visuels il apparaît clairement que la distance est également requise, et il est donc impératif de pouvoir déterminer les coordonnées 3D de la cible pour permettre à l'utilisateur de s'en saisir et ou de se déplacer vers elle.

Figure II-11 Caméra TOF (Fotonic E-Series)

Dans le cadre d'un système embarqué (ou *wearable device*), pouvant de plus être utilisé en extérieur, il existe assez peu de méthodes pour le calcul de cartes de profondeurs. Si certains capteurs lasers, radars ou acoustiques peuvent fournir des informations de profondeur, ils ne fournissent en revanche pas de flux vidéo permettant l'analyse de la scène ou la reconnaissance d'objets. Ils pourraient être couplés avec des caméras standards mais la synchronisation et la calibration est très délicate [Hussmann et al., 2008]. Parmi les solutions répondant à ces contraintes, les plus répandues sont les caméras binoculaires, reposant sur la stéréovision, ainsi que les caméras time-of-flight[1] (TOF). Ces dernières, beaucoup plus récentes, ont commencé à se démocratiser dans les années 2000. Elles reposent sur l'émission d'un signal lumineux infra-rouge pulsé et permettent de calculer le temps de parcours de la lumière réfléchie grâce à sa phase. Elles comprennent des LEDs ou des diodes lasers pour l'illumination autour du spectre infrarouge (voir Figure II-11), ainsi qu'un capteur CMOS pour l'acquisition de l'image (ceux-ci différant des capteurs CMOS classiques par leur capacité à mesurer au niveau de chaque pixel des différences d'intensité de l'ordre de la nanoseconde afin d'extraire la phase de la lumière pulsée incidente). Si elles se développent, ces caméras restent néanmoins peu courantes. Elles présentent certes quelques avantages comme un encombrement réduit, des vitesses de rafraîchissement élevées, et des tarifs abordables mais souffrent généralement de trop faibles résolutions[2], de bruit et d'une portée limitée, souvent inférieure à 5/10 mètres [Cui et al., 2010; Hahne and Alexa, 2008].

2.1.2 Stéréovision

Face aux contraintes des caméras TOF nous nous sommes tournés vers des méthodes de stéréovision pour la localisation des cibles en 3D. Les systèmes de vision stéréoscopiques, similaires dans leur principe à la perception humaine de la profondeur, sont basés sur l'utilisation de deux caméras légèrement espacées. Chaque point de l'environnement appartenant au champ de vision commun est observé selon deux points de vue différents et sa position peut ainsi être calculée par triangulation.

[1] Ou caméras temps de vol.
[2] Parmi les caméras TOF les plus connues, la SwissRanger 4000 de Mesa Imaging offre par exemple une résolution de 176×144 px, la CamBoard Nano, développée par PMD Vision, de 160×120px, tout comme la DepthSense311 de SoftKinetic ou les Fotonic E-Series.

Ce calcul nécessite la connaissance des paramètres intrinsèques et extrinsèques des caméras, estimés durant une phase préalable de calibration. Le modèle géométrique généralement utilisé pour représenter une caméra est le modèle sténopé représenté dans la Figure II-13. Celui-ci comprend un centre optique et un plan image. Les rayons lumineux issus d'un point P de l'espace convergent en ligne droite vers le centre optique et se projettent dans l'image en un point p, correspondant à l'intersection de cette droite avec le plan image. On appelle axe optique la perpendiculaire au plan image passant par le centre optique, et point central son intersection avec le plan image. Enfin la distance focale F correspond à la distance entre le centre optique et le plan image. Les paramètres intrinsèques de la caméra sont donc constitués par cette distance focale, le point central PC, ainsi que la taille de l'image.

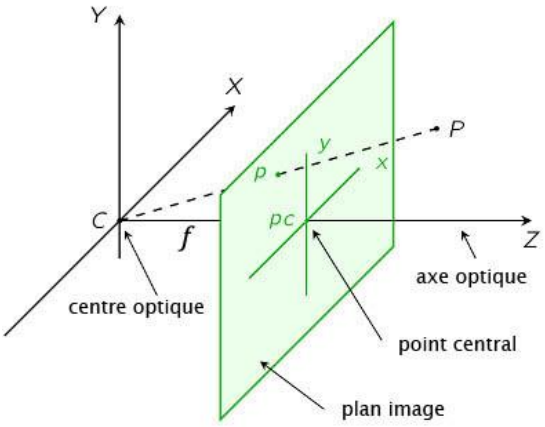


Figure II-13 Modèle géométrique de caméra

Les paramètres extrinsèques représentent quant à eux le changement de repère permettant de passer du référentiel d'une caméra à l'autre, ils comprennent 3 paramètres spécifiant la rotation entre les axes des deux repères, et 3 autres indiquant la translation entre leur origine. Une fois établies ces différentes valeurs propres à chaque caméra et à leur position relative il est possible, connaissant le projeté d'un point de l'espace dans le repère image de chacune, de calculer ses coordonnées, tel qu'illustré dans la Figure II-12.

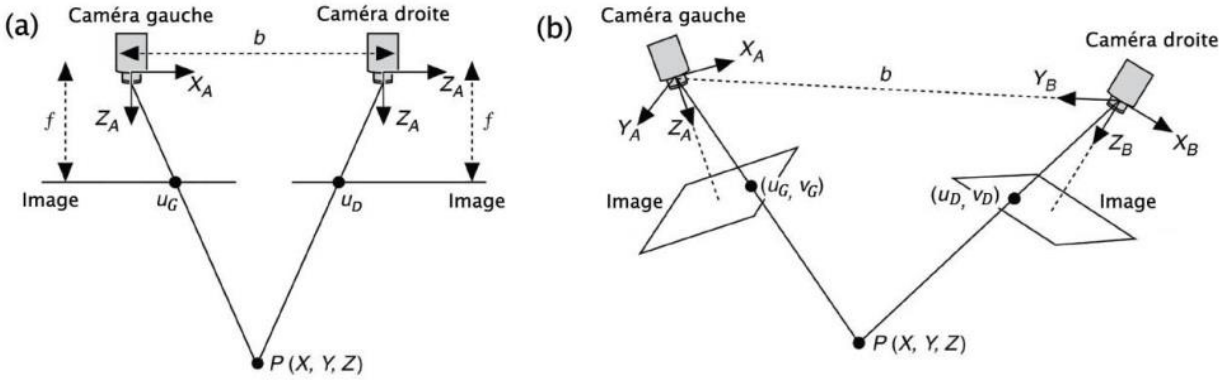


Figure II-12 Modèles géométriques de vision stéréoscopique

Dans le cas simple (à gauche de la figure), où les axes optiques des caméras sont parallèles et où les droites horizontales passant par le point central de chacune sont confondues, la profondeur D d'un point P est donnée par $D = f(b - disp)/disp$ où b

correspond à la distance entre les caméras, f leur focale, et $disp$ la disparité stéréo, c'est-à-dire la différence entre la projection du point dans l'image de gauche et de droite, soit $U_G - U_D$. Le cas à droite de la figure, plus complexe, repose néanmoins sur le même principe, mais impose quelques étapes supplémentaires de transformations géométriques pour tenir comptes des rotations respectives de chaque caméra, et d'une rectification des images de gauche et de droite pour les projeter sur un même plan (voir Figure II-14). L'optique de la plupart des caméras générant souvent différents types de déformations, la phase de rectification inclut généralement des corrections et des distorsions radiales, du décentrage des lentilles ou des erreurs de parallélisme [Tarel and Gagalowicz, 1995]. Ces différents paramètres sont également estimés durant la phase de calibration au moyen de mires géométriques.

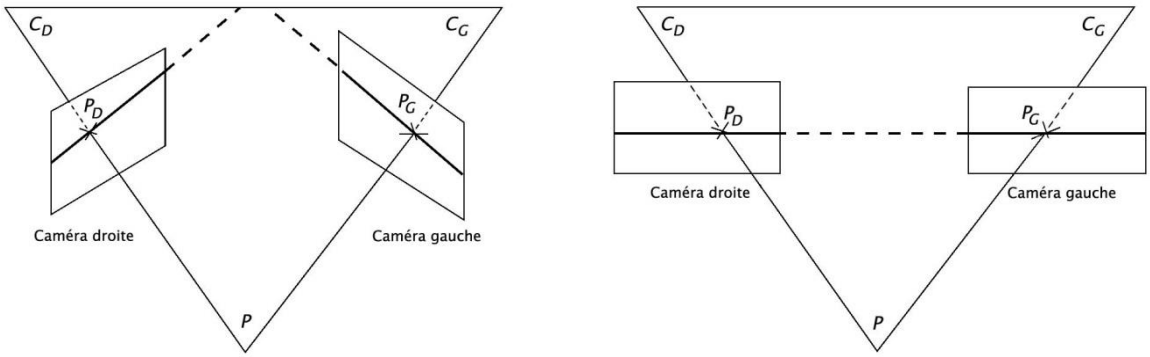


Figure II-14 Rectification stéréoscopique (projetions des images gauche et droite sur un même plan) ; à gauche avant, à droite après.

On distingue deux types de stéréovision, une dite dense, produisant une carte de profondeur donnant Z pour tout point de l'image, ou la stéréovision éparse pour laquelle on ne calcule la profondeur que sur des points caractéristiques (angles, coins, obstacles,…). Les deux cas nécessitent une phase d'appariement consistant à mettre en correspondance un point de l'image gauche avec son homologue dans l'image droite ou inversement. Pour cela la plupart des algorithmes proposés se basent soit sur une mesure de similarité telle que le coefficient de corrélation croisée appliqué directement à l'intensité des pixels dans une région autour du point candidat, soit sur des approches *feature-based* dérivant des caractéristiques locales de l'image (Dhond and Aggarwal, 1989, comme par exemple des histogrammes de gradients dans le cas des SIFT [Dhond and Aggarwal, 1989]. Cette deuxième méthode est plus robuste mais également plus coûteuse. Quelle que soit la solution adoptée, la recherche de la zone correspondante à un point donné n'est pas effectuée dans l'ensemble de l'autre image, mais simplement le long d'une ligne, en tirant parti des contraintes épipolaires illustrées dans la Figure II-15. Dans le cas d'images rectifiées, le calcul de cette droite épipolaire est encore plus simple, tout point se projetant

sur la même ligne de l'image gauche et droite. Une fois les points appareillés, il est finalement possible de calculer les coordonnées 3D à partir de leur disparité, comme expliqué précédemment.

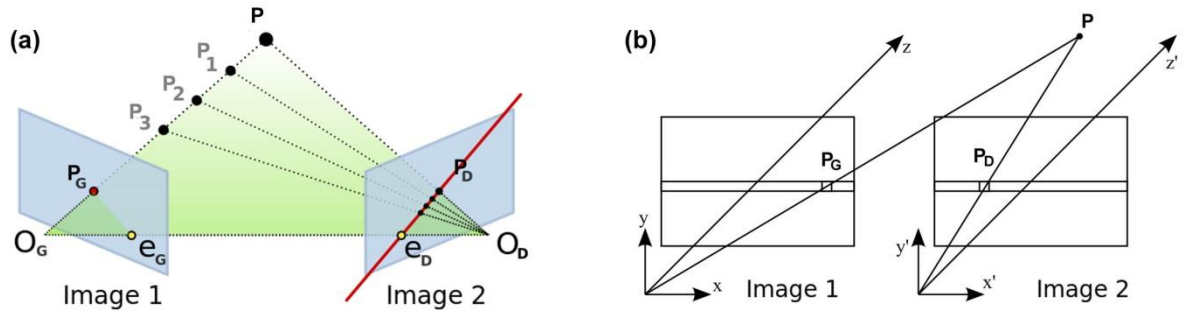


Figure II-15 Géométrie épipolaire : (a) l'ensemble des points entre le centre optique O_G et le point P se projettent dans l'image 1 au niveau de P_G et le long d'une droite correspondant à l'intersection entre le plan de l'image 2 et celui définit par les points$P$, O_G et O_D ; (b) en géométrie rectifiée, les projections d'un point de l'espace sur chaque image se trouvent nécessairement sur la même ligne.

2.1.3 Matériel

D'un point de vue pratique, la mise en place d'un système de stéréovision peut être effectuée à partir de deux caméras standard fixées sur un support rigide, dont les images capturées doivent être parfaitement synchronisées. Il est cependant nécessaire d'implémenter des procédures de calibration pour estimer les paramètres intrinsèques et extrinsèques mentionnés précédemment, ainsi que les propriétés de distorsion des optiques. Il faut enfin développer les algorithmes de rectification d'image, d'extraction de points d'intérêt et de mise en correspondance afin d'obtenir des cartes de disparité permettant de calculer la profondeur des points de l'image. Il existe également des solutions commerciales intégrant des caméras binoculaires relativement compactes. Celles-ci sont généralement pré-calibrées à l'achat, et disposent de librairies optimisées réalisant l'ensemble de la chaîne de traitement.

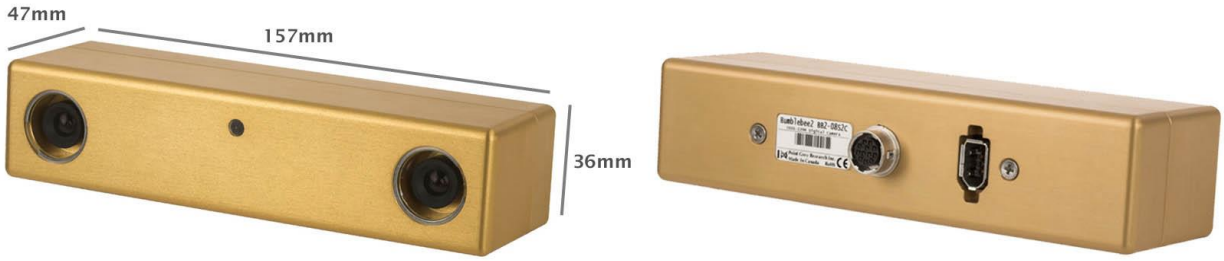


Figure II-16 Caméra BumbleBee (de face et de dos)

Après un premier dispositif réalisé à partir de webcams [Dramas, 2010], nous nous sommes finalement tournés vers un capteur spécialisé, offrant de meilleures performances en termes de précision et de vitesse d'exécution. Notre choix s'est porté sur la caméra BumbleBee2 (voir Figure II-16), commercialisée par Point Grey Research[1]. Elle comprend deux capteurs CCD d'une focale grand angle de 2,5 mm couvrant un angle de vue horizontal d'environ 100°. Les images, d'une résolution de 640x480 px, sont transmises par FireWire à 48 images par seconde. Les caméras sont livrées avec une suite logicielle complète. La libraire FlyCapture permet d'interfacer en C/C++ le contrôle des caméras et la réception des images (brutes ou rectifiées), alors qu'une autre librairie, nommée Triclops Stereo, offre les fonctionnalités liées à la stéréovision. L'algorithme implémenté repose sur un filtrage par Laplacien de Gaussienne permettant d'extraire les contours dans chacune des images, puis un appariement des points utilisant comme méthode de corrélation la somme des différences absolues. Ces étapes de traitement sont représentées dans la Figure II-17.

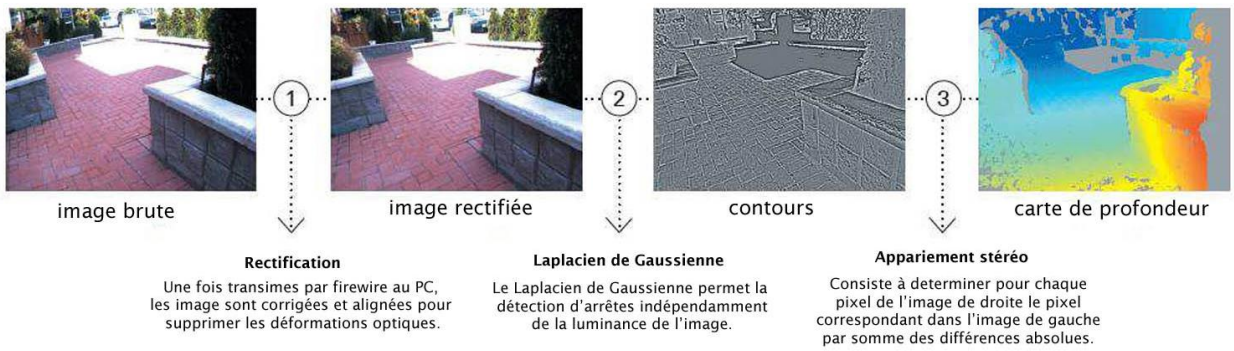


Figure II-17 Construction des cartes de profondeur

[1] http://ww2.ptgrey.com/stereo-vision/bumblebee-2

Cette caméra binoculaire a été fixée dans notre prototype sur un casque de vélo au moyen d'une rotule permettant d'orienter celle-ci aisément en fonction du port du casque. Sur celui-ci est également disposée une centrale inertielle comprenant gyroscopes, accéléromètres et magnétomètre, permettant de fournir les mouvements de la tête de l'utilisateur et donc ceux des caméras.

Figure II-18 Prototype Navig utilisant la caméra BumbleBee

2.1.4 Reconnaissance de formes

Il existe un grand nombre d'algorithmes permettant la reconnaissance visuelle de formes et d'objets préalablement appris. Néanmoins, l'utilisation de ceux-ci dans un dispositif d'aide aux non-voyants impose un certain nombre de contraintes, soulignées dans [Jafri et al., 2013] :

- Le système doit pouvoir fonctionner en temps réel pour que l'échange d'information avec l'utilisateur puisse être viable. Si plusieurs secondes sont nécessaires à l'identifiant d'un objet, l'utilisabilité du système restera très limitée.
- L'algorithme de reconnaissance doit pouvoir opérer correctement dans de nombreuses conditions, et donc supporter des variations d'illumination, de bruit, d'échelle, d'orientation, de point de vue, ou encore d'arrière-plan.
- Le dispositif doit être portable, c'est-à-dire de taille et de poids réduits, et ne pas entraver l'utilisateur dans ses tâches habituelles (comme les systèmes nécessitant l'utilisation des mains). Ce facteur introduit donc aussi des limites de charge computationnelle. Un algorithme nécessitant une grande puissance de calcul, par exemple un cluster ou un poste multiprocesseurs disposant d'importantes ressources matérielles ne sera évidemment pas adapté à une utilisation en situation de mobilité.

- Le système ne doit pas nécessiter d'adaptation spécifique de l'environnement ou d'infrastructure particulière pour fonctionner.
- Enfin, pour qu'il puisse être adopté par un nombre important de non-voyants, il est important de veiller à ce que le coût du dispositif reste modéré, et donc s'abstenir d'utiliser des équipements ou technologies trop chers.

Pour ces différentes raisons, notre choix de l'algorithme de reconnaissance et de localisation des cibles visuelles s'est porté sur le moteur Spikenet, une méthode de reconnaissance de formes offrant une grande tolérance au bruit, au flou et aux conditions lumineuses, ainsi que des temps de reconnaissance particulièrement faibles. Elle a été développée au sein du laboratoire Cerveau et Cognition de Toulouse, par Simon Thorpe, Ruffin VanRullen et Arnaud Delorme, en s'inspirant des traitements effectués par le système visuel humain [Delorme et al., 2001, 1999; Delorme and Thorpe, 2003; Thorpe et al., 2001, 2004, 2000; VanRullen et al., 2005, 1998; VanRullen and Thorpe, 2002]. Cette modélisation des traitements neuronaux sous forme de réseaux impulsionnels asynchrones s'étant avérée particulièrement efficace pour la reconnaissance d'objets, ou la détection de visages, elle a donné lieu à la création de la société Spikenet Technology, qui a enrichi l'algorithme originel de nombreuses optimisations et commercialise cette libraire dans des solutions industrielles à destination de secteurs aussi variés que l'indexation de vidéos, la surveillance des tables de jeux dans les casinos, l'analyse d'images pour le trafic routier, l'estimation de foules, ou encore les domaines de la sécurité (vidéo protection, détection d'intrusions, etc.).

Les aspects computationnels de l'algorithme Spikenet (comme la nature même des traitements réalisés, et leurs propriétés) seront développés dans le chapitre III. Ici, nous n'aborderons que les aspects fonctionnels, c'est-à-dire les questions relatives à l'apprentissage des modèles de cibles visuelles, à leur activation ainsi que l'utilisation de la reconnaissance de formes au sein du dispositif de suppléance Navig. Soulignons que si cet algorithme a été retenu dans le projet Navig pour sa rapidité d'exécution, compatible avec une utilisation temps-réel par un système de suppléance, et pour sa tolérance aux déformations, au bruit, ou aux conditions d'illumination, le principe même du système n'est pas intrinsèquement lié à cet algorithme en particulier. Les différentes fonctions du module de vision au sein du dispositif pourraient tout aussi bien être assurées par une autre méthode de reconnaissance de formes, pourvu qu'elle satisfasse aux exigences en fiabilité et en précision de localisation, ainsi qu'en temps de traitement.

Précisons pour finir que la recherche de cibles visuelles n'est réalisée que sur une des deux images rectifiées des caméras (par convention l'image de droite est utilisée comme référence), puis pour chaque détection, la librairie Triclops permet -par les méthodes de

stéréovision expliquées plus tôt- d'obtenir les coordonnées métriques 3D de l'objet à partir de sa position dans l'image.

2.2 Localisation d'objets

Ce mode de fonctionnement du système repose sur l'utilisation en boucle fermée de la vision artificielle et du moteur de sonification spatiale [Dramas et al., 2010], illustré dans la Figure II-19. Il permet au non-voyant de localiser grâce aux sons virtuels 3D la position d'un objet ou d'une cible pour éventuellement s'en saisir si elle se trouve dans l'espace péripersonnel, où se diriger vers elle si elle se situe à distance.

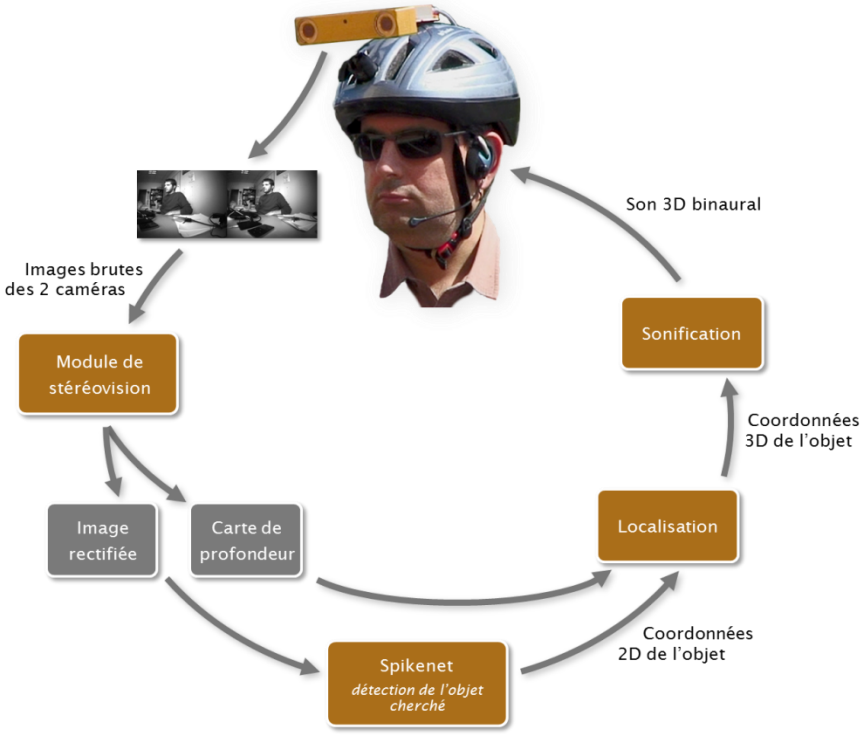


Figure II-19 Schéma de l'architecture du système de localisation d'objet. Les images acquises par les caméras sont transmises au module de stéréovision produisant en sortie une image rectifiée (corrigeant les déformations optiques), et une carte de profondeur (donnant la distance de chacun des pixels de l'image rectifiée). La détection de l'objet est réalisée par l'algorithme Spikenet, et permet, à partir de ses coordonnées dans l'image et de la carte de profondeur, de calculer la position 3D de la cible dans le repère caméras. Un son virtuel spatialisé est finalement généré aux coordonnées trouvées, et transmis à l'utilisateur.

Le mode « localisation d'objet » peut être activé dans 3 cas d'usage. Premièrement, lorsque l'utilisateur sollicite explicitement cette fonction pour la recherche d'un objet particulier (comme un téléphone dans la Figure II-20), en accompagnant sa requête du nom de l'objet souhaité. Celui-ci, interprété par le contrôleur de dialogue grâce au module de reconnaissance vocale, est comparé aux identifiants de la base d'objets préalablement appris, et s'il existe, son ou ses modèles sont alors chargés dans le moteur de reconnaissance Spikenet. Le deuxième cas d'utilisation intervient dans le mode de navigation, en fin de parcours, lorsque la destination correspond à une cible visuelle. Ainsi, si l'utilisateur désire se rendre au laboratoire d'informatique de Toulouse, il sera guidé depuis sa position courante grâce au GPS et au système d'information géographique, comme décrit dans la section précédente, et lorsqu'il atteindra la destination finale, le système activera la détection visuelle de la façade du bâtiment puis de la porte principale en fonction de la distance, pour le guider de façon précise jusqu'à son but, comme illustré dans la Figure II-21. Le dernier usage correspond à l'activation temporaire de la localisation d'objets d'intérêt pour l'utilisateur au cours du mode classique de navigation (se basant donc sur le positionnement GPS). Ces objets sont sélectionnés par le contrôleur de dialogue en fonction de la position de l'utilisateur et leurs coordonnées géolocalisées dans le SIG. Il est par exemple possible de détecter les bandes blanches des passages piétons grâce à des modèles génériques assez robustes. Lorsqu'au cours d'un trajet le non-voyant aura à franchir une rue, le système pourra donc interrompre le guidage par balises sonores positionnées aux points d'itinéraire, pour plutôt préférer un guidage précis grâce à la localisation visuelle d'objets. Une fois la traversée effectuée, le guidage standard pourra reprendre la main.

Figure II-20 Localisation et préhension d'objet. Le téléphone recherché est reconnu par le module de vision, et un son 3D est synthétisé à la position calculée par stéréovision.

Que la demande de recherche d'un objet provienne du contrôleur de dialogue au cours ou en fin d'un trajet, ou de l'utilisateur lui-même, à tout moment, les mécanismes impliqués restent identiques. Différents modèles de l'objet en question sont alors chargés dans le noyau Spikenet afin de couvrir différents angles de vue, orientations, échelles de celui-ci, ainsi que différentes variations individuelles dans le cas de modèles génériques représentant une catégorie d'objets. Parmi les possibles détections sur une image donnée, celle au score le plus haut est conservée, ses coordonnées 3D sont calculées par stéréovision puis retournées à l'agent IHMS, en charge de la sonification, au moyen d'un message Ivy contenant le nom de la cible et sa position. Un son 3D à cette position est alors généré en simulant artificiellement les phénomènes acoustiques de perception auditive spatiale grâce aux HTRF (se reporter à la section II.1.4). Il peut s'agir d'un son simple ou de parole selon les préférences de l'utilisateur et le contexte. La recherche de l'objet par le module de vision est ensuite répétée sur toutes les images ultérieures reçues des caméras jusqu'à interruption par l'utilisateur ou le contrôleur de dialogue. La vitesse de cette boucle dépend bien sûr du rafraichissement des caméras, fixant une limite maximale à 48Hz, mais également du temps de traitement requis par l'algorithme de reconnaissance. Celui-ci est fonction du nombre de modèles activés de leur taille, ainsi que de celles de l'image. En pratique il peut varier de seulement 50 ms à parfois 400 ou 500 ms si de très nombreux modèles sont recherchés à de multiples échelles. Cependant dès la première détection d'un objet, les sons générés et restitués à l'utilisateur pour le localiser seront répétés à intervalle fixe. Celui-ci peut être paramétré, mais différentes expérimentations de préhension de cibles ou de déplacement nous ont montré qu'une fréquence de répétition autour de 15hz offrait de bons résultats, à savoir une bonne précision de localisation sans pour autant surcharger l'utilisateur de sons trop invasifs [Macé et al., 2012].

Dans le cas idéal où la vitesse des traitements visuels est supérieure à celle de la sonification, il suffit au module IHMS d'utiliser les dernières coordonnées reçues. Dans le cas contraire, l'utilisation de la centrale inertielle dont le casque est équipé permet de compenser les mouvements de la tête effectués depuis la dernière détection. La méthode utilisée consiste, à chaque réception d'un message contenant les coordonnées de la cible, à stocker celles-ci. Elles seront ainsi associées à la position du casque à cet instant donné (c'est-à-dire les valeurs de rotation sur 3 axes). Ensuite, à chaque émission d'un nouveau son spatialisé, les dernières coordonnées de l'objet sont corrigées en fonction du changement d'orientation de la tête depuis la détection. De cette façon nous pouvons conserver une localisation cohérente de la cible malgré une fréquence de détection visuelle trop faible, mais également en l'absence de détection. Cela s'avère très utile et efficace lorsque l'utilisateur balaye « visuellement » la scène. En effet si un objet est par exemple détecté à droite du champ de vision et que l'utilisateur tourne ensuite la tête vers la gauche, il sortira du champ des caméras mais les sons virtuels continueront néanmoins d'être présentés à la

position correcte de l'objet (si celui-ci est fixe). Ce mécanisme est d'autant plus nécessaire que les sujets ont une forte tendance à scanner l'environnement par des mouvements de la tête pour améliorer la perception auditive spatiale.

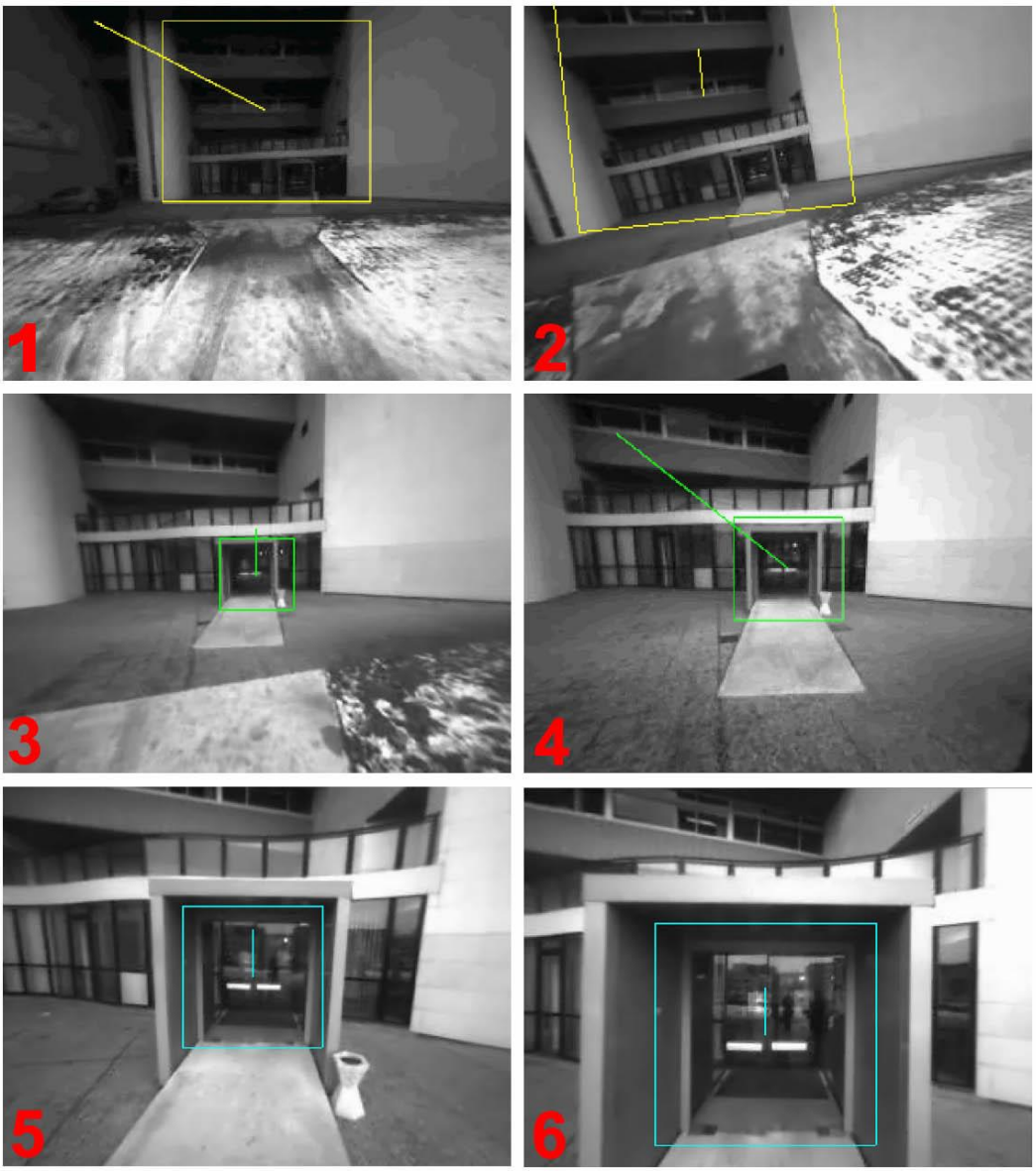


Figure II-21 Navigation vers une destination assistée par la vision artificielle. Plusieurs modèles visuels permettent de détecter l'entrée du laboratoire à différentes échelles, orientations et détails afin que l'utilisateur puisse se diriger vers celle-ci. Les chiffres en rouge indiquent l'ordre chronologique des images acquises au cours de ce déplacement et les rectangles en couleur correspondent aux détections effectuées par Spikenet.

2.3 Positionnement utilisateur

Les systèmes d'aide à l'orientation (EOA[1]) reposent en grande majorité sur une architecture standard constituée de 3 composants :

- Un module de positionnement, permettant de localiser l'utilisateur.
- Un module cartographique c'est à dire un système d'information géographique comprenant des fonctionnalités de calcul et de suivi d'itinéraire.
- Enfin un module de restitution de l'information, pouvant utiliser des instructions de guidage par synthèse vocale, par dispositif tactile, ou encore des balises sonores spatialisées.

Jusqu'à présent ces EOA sont principalement basées sur le GPS (comme BrailleNote, Angeo, Trekker), et dans la plupart des cas, leur utilisation a été limitée par une mauvaise précision de positionnement (souvent supérieure à 10m, en particulier dans les zones urbaines), qui peut mener à des situations dangereuses (traverser en dehors des zones protégées) ou à des erreurs (emprunter la mauvaise rue). Il apparaît donc clairement, voir par exemple [Gaunet and Briffault, 2005], qu'une localisation précise (c'est-à-dire inférieure à 5 mètres), est cruciale pour l'orientation dans les points délicats tels que les intersections, les passages piétons, ou pour déterminer de quel côté de la rue l'utilisateur se trouve.

Nous détaillerons donc ici une nouvelle méthode de positionnement permettant de compenser les limites du GPS par l'utilisation de la vision artificielle et la détection d'amers visuels[2] (indépendamment de la reconnaissance de cibles d'intérêt sonifiées à l'utilisateur). Cette approche novatrice s'inspire de fonctions visuelles et cognitives supportant la navigation. La reconnaissance d'éléments ou de lieux tels que le clocher d'une église, un magasin, ou autre bâtiment dont l'emplacement est connu, permet en effet à une personne voyante d'estimer sa position, de s'orienter dans son environnement et d'effectuer des inférences spatiales quant à celui-ci. Les mécanismes sous-tendant cette faculté reposent sur trois éléments :

- L'identification visuelle d'une cible
- La connaissance de la position absolue de celle-ci (dans un référentiel fixe de type cartographique)
- L'estimation de sa position relative par rapport à soi (distance et azimut)

[1] Abréviation d'*Electronic Orientation Aids.*
[2] Un amer est un point de repère fixe et identifiable sans ambiguïté utilisé pour la navigation.

Notre proposition consiste donc à les substituer par l'apprentissage et la reconnaissance de cibles visuelles caractéristiques, devant être géolocalisées dans un système d'information géographique, et dont la stéréovision permet de calculer la localisation par rapport à l'utilisateur.

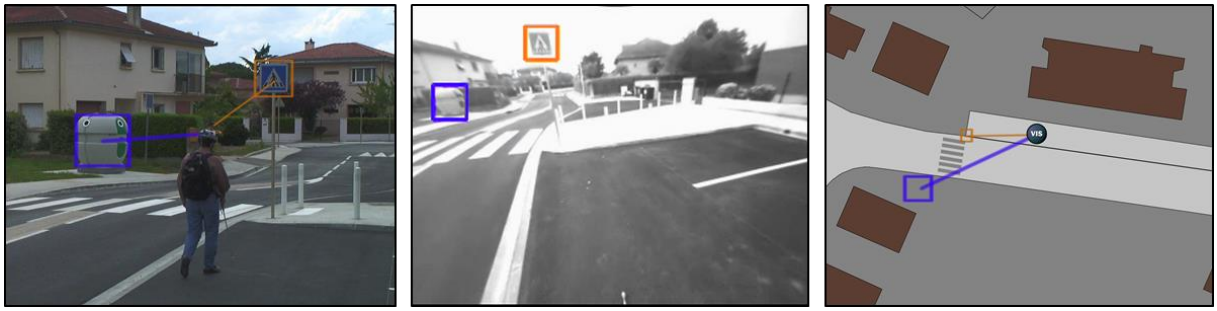


Figure II-22 Positionnement utilisateur basé sur la détection d'amers visuels aux coordonnées connues. Vue extérieure à gauche, image des caméras embarquées au centre droite, et vue de la carte à droite.

2.3.1 Technologies traditionnelles de positionnement

Positionnement par satellites

Les récepteurs GPS constituent les méthodes de positionnement les plus répandues. Leur précision varie généralement entre 5 et 60 mètres. Des améliorations technologiques récentes, telles que l'Assisted GPS[1], ont certes permis de réduire l'erreur nominale de près de 100 mètres à moins de 10, mais elle demeure néanmoins toujours trop élevée pour les besoin de locomotion et d'orientation d'un piéton non-voyant.

Cette erreur est le résultat de plusieurs facteurs combinés, le plus important étant l'occultation des signaux ou leur réflexion sur des bâtiments. En milieu urbain ce problème est très fréquent, et les multi-trajets dus aux rebonds du signal entraînent un temps supplémentaire à la réception pouvant induire jusqu'à 50 mètres d'erreur. S'ajoutent à celle-ci l'effet des conditions atmosphériques, des arrondis de calcul, des impressions des horloges, ou encore de la différence entre la position réelle d'un satellite et son orbite

[1] L'Assisted GPS, au moyen d'une connexion internet, permet le téléchargement de la table d'éphéméride des satellites nécessaire au calcul de la position. Celle-ci est ci n'a une durée de vie que de 4 heures, et doit donc habituellement être reçue à chaque lancement des satellites, dont la vitesse de transmission est lente, l'A-GPS permet donc un fix plus rapide. Il offre aussi la possibilité d'envoyer les données reçue à un serveur et d'y déporter le calcul de la position. Celui-ci, disposant de ressources de calcul plus importantes, et d'une bonne couverture satellite peux ainsi comparer les signaux fragmentés reçus de l'utilisateur avec ceux d'une antenne relais pour améliorer le positionnement.

théorique. Chacun de ces facteurs peut entraîner jusqu'à 2 ou 3 mètres d'erreurs supplémentaires.

Une autre technologie GPS s'est répandue dans la navigation maritime au cours des années 90. Baptisée D-GPS (GPS Différentiel), elle repose sur un réseau de stations de référence dont les coordonnées exactes sont connues, qui calculent et transmettent l'écart entre ces positions et celles fournies par le GPS. Les corrections obtenues peuvent être appliquées au receveur D-GPS, permettant de réduire l'erreur de positionnement commune aux deux GPS. Si dans des conditions dégagées ce système permet d'obtenir une précision de 1 à 3 mètres, il ne permet cependant pas de corriger les problèmes de rebond (ou multi-trajet) fréquents en environnement urbain. L'équipement nécessaire est de plus généralement lourd et encombrant, et l'utilisation des services D-GPS nécessite un abonnement payant auprès des agences gouvernementales ou des sociétés privées implantant les stations de références. Au nombre d'une trentaine en France, majoritairement concentrées sur le littoral ou en montagne, leur couverture est pour finir relativement faible. Pour ces différentes raisons, le D-GPS ne constitue donc pas une solution viable pour la navigation piétonne.

Navigation à l'estime

Pour compenser les imprécision du GPS et les pertes temporaires de signal, plusieurs systèmes ont proposé de coupler le positionnement par satellites à des méthodes de navigation à l'estime, ou *dead-recknoning* [Beauregard and Haas, 2006; Jimenez et al., 2009]. Celles-ci consistent à déterminer le déplacement relatif depuis la dernière position connue. Si ce type d'approche offre souvent de très bons résultats dans le cadre de véhicules ou de robots, le problème est beaucoup plus complexe pour un piéton. En effet les techniques de *dead-reckoning* reposent généralement sur des modèles probabilistes du déplacement, et si ceux d'une voiture, par exemple, répondent à des lois assez simples, tant au niveau de la trajectoire que de la dynamique, les mouvements humains sont eux, beaucoup plus durs à prédire. De plus les véhicules comme les robots peuvent être équipés de capteurs procurant de façon précise les différents paramètres de leur déplacement. Connaissant la dimension des roues d'une voiture et leur empattement, il est ainsi possible d'obtenir sa trajectoire en mesurant la rotation des roues et de la colonne de direction. Pour les humains la plupart des systèmes se basent sur des centrales inertielles. Les accéléromètres qu'ils contiennent peuvent être utilisés comme podomètres, alors que le compas magnétique fournit l'orientation. Ces mesures étant beaucoup moins précises (les compas magnétiques sont soumis à de nombreuses perturbations extérieures, et la longueur moyenne des pas n'est pas toujours fiable), les accéléromètres souffrent donc de

performances relativement modestes, et ne peuvent être utilisés que sur des périodes courtes, les biais et erreurs s'accumulant au cours de l'intégration temporelle.

Chez un piéton, ces centrales inertielles peuvent-être placées sur le pied, [Fischer et al., 2008; Jimenez et al., 2009], sur un casque [Beauregard, 2006], ou encore à la hanche [Stirling et al., 2003]. Le positionnement sur le pied offre certains avantages, une détection des pas plus robuste, et une correction de la dérive en réinitialisant les vecteurs d'accélération grâce à une méthode appelée Zero Velocity Update [Feliz Alonso et al., 2009], néanmoins le port sur la ceinture est beaucoup plus commode, et reste une des solutions offrant les meilleures performances [Stirling et al., 2003].

2.3.2 Vers un positionnement hybride utilisant la vision artificielle

Comme l'ont montré de nombreuses études, la clé pour une localisation précise d'un piéton consiste en la combinaison de différentes stratégies de positionnement. Nous avons proposé dans le cadre du système Navig une solution novatrice n'ayant à notre connaissance jamais été mise en place jusqu'à présent, consistant en une fusion de trois méthodes, à savoir le positionnement par satellites, la navigation à l'estime, et la détection de cibles visuelles géolocalisées [Brilhault et al., 2011]. Ces amers visuels détectés au cours d'un trajet, telles que des enseignes de magasins ou tous autres motifs caractéristiques, permettent ainsi de compenser des signaux GPS dégradés ou totalement masqués en intérieur ou dans des passages couverts.

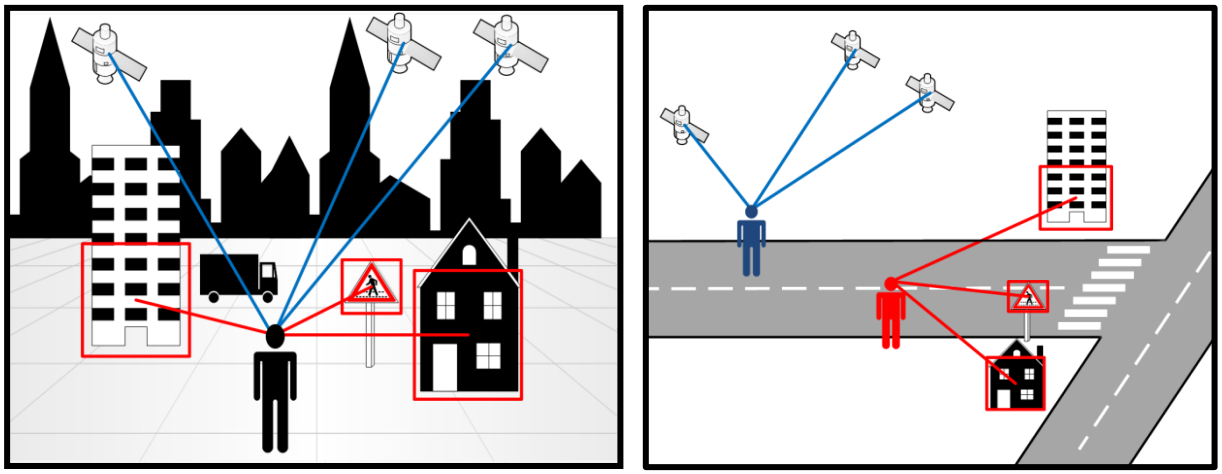


Figure II-23 Positionnement par satellites et vision artificielle

Notre méthode s'appuie sur un capteur GPS, une ou plusieurs centrales inertielles, ainsi qu'un système de vision artificielle utilisant des caméras stéréoscopiques. Notre dernier prototype intègre deux centrales inertielles, une positionnée sur un casque, fournissant

l'orientation des caméras, et une seconde portée à la ceinture, utilisée comme podomètre et compas magnétique. Le système de vision, présenté précédemment, repose sur une caméra stéréo BumbleBee ainsi que sur l'algorithme Spikenet. Enfin le positionnement satellite peut être assuré par un récepteur GPS indépendant standard ou par celui d'un téléphone portable, communiquant en Bluetooth avec le reste du système. Nous avons aussi utilisé le boîtier Angéo, présenté dans la Figure II-4 (à la page 103). Il s'agit d'un boîtier développé par l'entreprise NAVOCAP de 300 g mesurant 112 x 72 x 18 mm, qui intègre un GPS HighSense et l'Assisted-GPS grâce à sa prise en charge des communications GSM/GPRS/EDGE quadband. Il contient également gyromètres, accéléromètres, et magnétomètres 3D, et repose sur un processeur Freescale ARM9 à 400 Mhz ainsi que sur 8Go de stockage. Un algorithme propriétaire breveté permet de fournir en sortie des coordonnées GPS consolidées, plus robustes en théorie que celles des récepteurs standards. Néanmoins la mise en place de ce dispositif par la société NAVOCAP, partenaire du consortium Navig, s'est faite en parallèle du développement de notre prototype, et les différentes versions qu'ils nous ont mises à disposition au cours des 3 ans qu'ont duré ce projet, n'étaient pas encore finalisées, et offraient des performances souvent aléatoires. Nous nous sommes donc tournés vers des récepteurs GPS classiques dans la plupart de nos expérimentations.

2.3.3 Gestion des modèles visuels

La présence et la position des cibles détectées dans le mode « positionnement utilisateur », contrairement au mode « localisation d'objet », ne sont pas restituées à l'utilisateur mais seulement utilisées pour être fusionnées avec le GPS afin d'améliorer la précision de positionnement. Ce n'est donc pas à l'utilisateur de déterminer quelles cibles doivent être activées, comme dans le mode de localisation d'objets, mais au système, qui chargera automatiquement en fonction de la position courante le lot de modèles à rechercher. Cette sélection de cibles potentielles est cruciale, car si tous les amers visuels d'une ville devaient être chargés simultanément, leur nombre risquerait d'être trop important pour des détections en temps réel. De plus, le fait de se limiter aux cibles environnantes permet également de réduire les risques de fausses détections (à la fois les fausses alarmes classiques, en l'absence de la cible, mais également lorsque plusieurs objets identiques existent à différentes coordonnées GPS, comme un panneau stop, ou le logo d'une chaîne de magasins par exemple).

Ces amers étant géolocalisés dans le SIG sous la catégorie Points Visuels, le module vision doit lui transmettre une requête à chaque nouvelle position reçue, afin de recevoir la liste des identifiants des PVs à proximités. Les modèles Spikenet correspondant à ceux-ci pourront ainsi être activés dans le moteur de reconnaissance. Deux méthodes permettent de

rapatrier cette liste de points depuis le SIG. La première, plus intuitive, consiste à simplement retenir l'ensemble des PVs dont les coordonnées se trouvent à une distance donnée de la position courante (en pratique nous utilisons souvent un rayon proche de 30 mètres pour compenser les erreurs possibles du GPS). La deuxième stratégie consiste à utiliser un champ facultatif des éléments du SIG indiquant la portée d'un objet. Celui-ci est par exemple utilisé pour la sonification des points d'intérêt, un arrêt de bus étant signalé seulement lorsque l'utilisateur se trouve à quelques mètres de celui-ci, alors que d'autres éléments tels que les musées pourront être indiqués même à 50 mètres. Dans le cas des points visuels, ce champ caractérise la distance à laquelle l'objet est censé pouvoir être détecté. Ainsi, s'il est nécessaire de se trouver très proche d'un panneau de signalisation pour reconnaître celui-ci (au-delà, étant donné sa taille et l'angle de vision des caméras, il sera trop petit dans l'image pour être correctement détecté), une église, ou une façade de bâtiment pourra à l'inverse être visible à une distance importante[1]. La primitive d'accès au SIG prenant en compte la portée consiste à rechercher les points pour lesquels la position de l'utilisateur se trouve à l'intérieur du cercle autour de ceux-ci d'un rayon égal à leur portée, tel qu'illustré dans la Figure II-24.

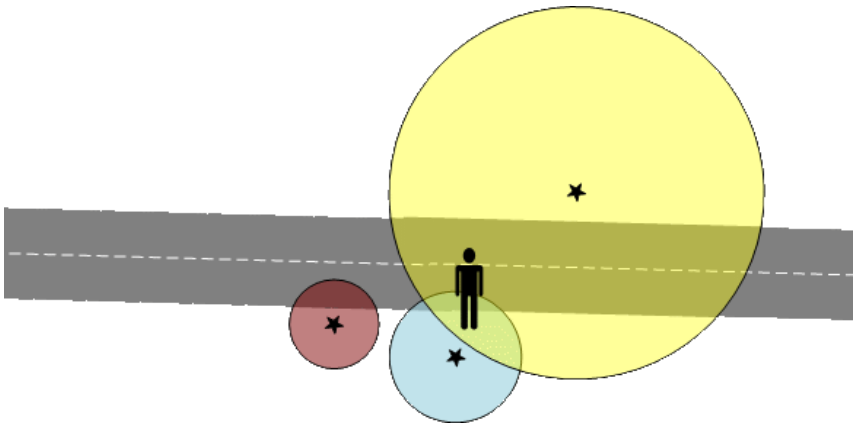


Figure II-24 Illustration de la portée des points visuels au niveau du SIG. Dans cet exemple les deux points retournés seront ceux matérialisés par le disque jaune est bleu.

Du point de vue de l'implémentation, si le chargement ou la suppression de modèles au niveau du moteur Spikenet a un coût relativement faible au niveau du temps de traitement, celui-ci n'en reste néanmoins pas négligeable, à plus forte raison si ces opérations doivent être effectuées toutes les secondes, lors de chaque mise à jour de la

[1] Cette portée a généralement été renseignée manuellement pour certaines cibles dont la valeur par défaut semblait trop faible, mais il est possible d'imaginer calculer cette valeur de façon automatique connaissant la taille physique de l'objet, où la distance à laquelle ont été prises les images utilisées pour son apprentissage par Spikenet.

position. Lors du chargement d'un nouveau modèle, celui-ci se voit affecter un identifiant temporaire unique (partant de 1 et augmentant à chaque modèle), libéré à la suppression de celui-ci. Les différentes primitives de gestion des modèles offertes par Spikenet sont résumées ci-dessous. Lorsqu'elles concernent un modèle, cet identifiant temporaire doit être fourni :

- Chargement (retourne son identifiant)
- Suppression d'un ou de tous les modèles
- Activation ou désactivation d'un ou de tous les modèles
- Obtention du nombre de modèles actuellement chargés
- Obtention de la liste des identifiants des modèles actuellement chargés
- Obtention des informations relatives à un modèle (son nom, description, et les différents paramètres de détection).
- Modification des informations relatives à un modèle

La gestion des modèles actifs au cours d'un trajet se révèlerait donc assez coûteuse en se limitant à ces primitives d'accès. Elle imposerait de supprimer l'ensemble des modèles à chaque rafraîchissement de la position, puis de charger tous ceux reçus du SIG pour celle-ci. Il serait également possible de récupérer l'ensemble des identifiants actuels, puis leurs noms (un par un), devant ensuite être comparés avec la liste des points courants, pour savoir s'ils doivent être maintenus ou supprimés, en marquant au passage chacun de ces points afin de déterminer les modèles qui, au final, doivent être chargés. Comme en pratique un faible nombre de modèles doit être changé à chaque seconde étant donné la vitesse de déplacement d'un piéton, ces solutions apparaissent clairement non adaptées. Nous avons donc intégré au module vision une surcouche permettant une gestion optimale des modèles. Celle-ci inclut plusieurs tables de hachages contenant les modèles chargés et les modèles actifs, associant leur identifiant temporaire et leur nom (il s'agit plus précisément de multimap, plusieurs modèles pouvant être associés à un nom unique de cible). De cette manière il est possible, à chaque réception d'une liste de points, de comparer ceux-ci avec les modèles actuellement actifs sans appel au noyau Spikenet et avec un nombre minimal d'opérations, pour ensuite n'appliquer que les modifications requises au niveau du moteur de reconnaissance. Cette surcouche permet également d'assurer la commutation de contexte entre positionnement utilisateur et localisation d'objets. En effet, lorsque le guidage visuel est déclenché à la demande de l'utilisateur ou du contrôleur de dialogue, les amers visuels doivent être désactivés pour offrir une boucle d'interaction la plus rapide possible.

2.4 Moteur de fusion

Le moteur de fusion, implémenté par Jiri Jiri Borovec, constitue le module du système NAVIG dédié au calcul d'une position consolidée à partir des différentes sources disponibles. Comme nous l'avons mentionné lors de la présentation du système, les différents agents communiquent au moyen de messages textes échangés via le bus logiciel Ivy. L'agent Fusion reçoit par ce biais les coordonnées GPS brutes fournies par le capteur GPS (ou le boitier Angéo), les données en provenance des deux centrales inertielles, ainsi que les détections du module vision, comprenant l'identifiant de la cible reconnue (permettant de récupérer ses coordonnées dans le SIG), et sa position cartésienne dans le référentiel caméra. Ces différentes interactions sont représentées dans la Figure II-25.

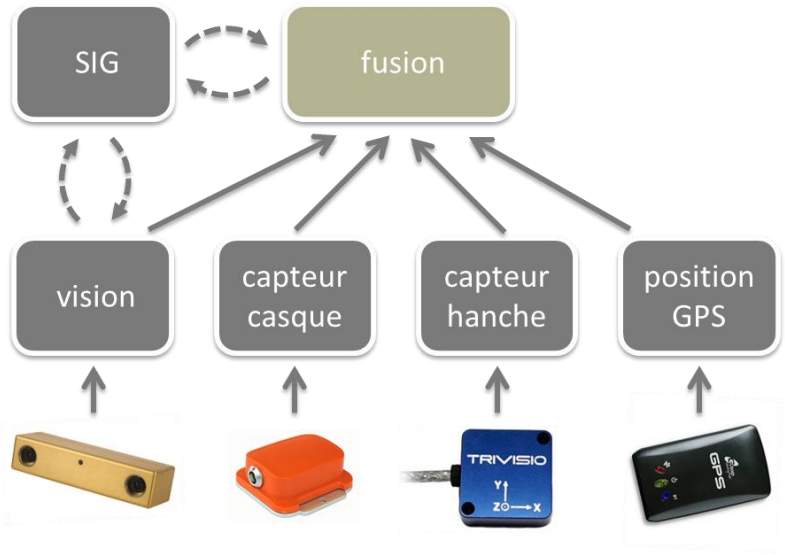


Figure II-25 Interaction du module de fusion au sein du système Navig

A partir de ces différents capteurs, l'algorithme de fusion permet d'estimer la position de l'utilisateur selon un processus se décomposant en trois phases : les prétraitements, l'autocorrection, et le positionnement final. L'architecture complète de ce module est représentée dans la Figure II-26, et ses différents composants seront détaillés au cours de cette section.

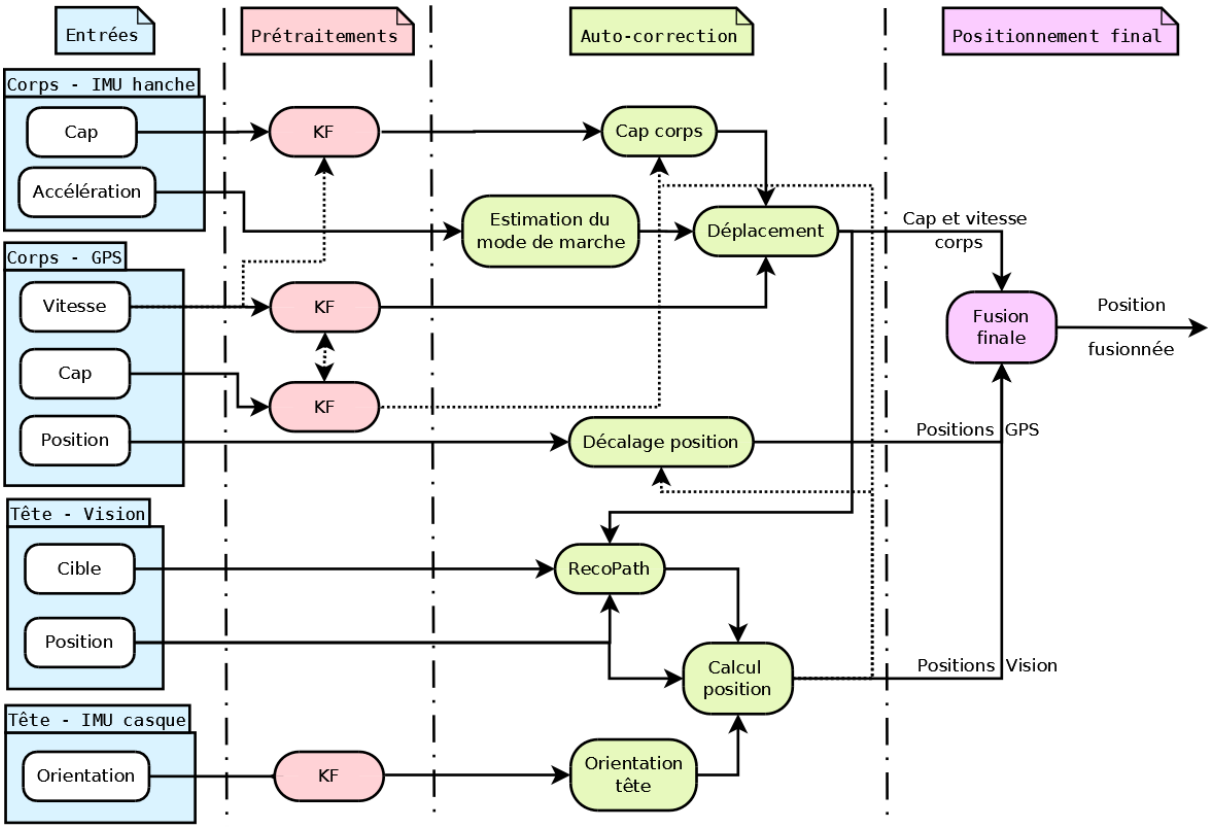


Figure II-26 Architecture du module de fusion. Les prétraitements consistent en un filtrage du bruit et des discontinuités. L'autocorrection inclut des processus plus complexes comme l'élimination des fausses détections visuelles, l'estimation de la marche, la correction de l'orientation,... Enfin la fusion finale consiste à déterminer la position de l'utilisateur à partir des informations extraites dans les phases précédentes. (Les flèches pleines représentent les flots de données, alors que celles en pointillés symbolisent les influences entre sous-modules, par exemple les entrées modifiant les paramètres des filtres de Kalman)

2.4.1 Position basée sur la vision

Lorsqu'un point visuel[1] est détecté au cours de l'itinéraire, l'agent Vision émet un message contenant l'identifiant de celui-ci, ainsi que ses coordonnées cartésiennes (x, y, z) dans le repère caméra, comme nous l'avons vu précédemment. L'estimation de la pose[2] d'une caméra monoculaire dont le mouvement est contraint sur le plan horizontal nécessite, tel que démontré dans [Burschka and Hager, 2003], trois cibles fixes dont les coordonnées

[1] Au sens défini dans la terminologie du SIG présenté dans la section II.1.6, à savoir un élément de l'environnement préalablement appris, pouvant être détecté par le moteur Spikenet, et dont les coordonnées sont connues.
[2] C'est-à-dire sa position et son orientation.

dans un système de référence global sont connues. Dans notre cas, comme nous utilisons des caméras stéréoscopiques et que la centrale inertielle fournissant leur orientation intègre un magnétomètre donnant celle-ci relativement au nord magnétique, une seule cible est requise pour calculer la latitude et la longitude de l'utilisateur.

L'orientation des caméras est fournie sous la forme d'angle d'Euler (*yaw, pitch,* et *roll*). Tel que montré dans l'équation ci-dessous, il est possible, en multipliant les coordonnées (x, y, z) de la cible dans le référentiel caméras par les matrices de rotation pour chacun des angles, d'obtenir ses coordonnées (x', y', z') dans le référentiel global. L'altitude z' ne sera plus prise en compte par la suite car nous considérerons que la position au sol de l'utilisateur.

$$\begin{bmatrix} x' \\ y' \\ z' \end{bmatrix} = \begin{bmatrix} x \\ y \\ z \end{bmatrix} . \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(yaw) & \sin(yaw) \\ 0 & \sin(yaw) & \cos(yaw) \end{bmatrix} . \begin{bmatrix} \cos(pitch) & 0 & -\sin(pitch) \\ 0 & 1 & 0 \\ \sin(pitch) & 0 & \cos(pitch) \end{bmatrix} . \begin{bmatrix} \cos(roll) & \sin(roll) & 0 \\ \sin(roll) & \cos(roll) & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Une requête au SIG, contenant la géolocalisation de tous les points visuels, permet ensuite d'obtenir la latitude et la longitude de la cible détectée dans le système géodésique mondial (WGS 84). A partir de celles-ci, de la distance de l'objet, et de l'inverse de l'orientation par rapport au nord magnétique nous pouvons finalement calculer les coordonnées de l'utilisateur dans le système WGS84.

Cette localisation est néanmoins soumise à certains facteurs d'imprécision. Tout d'abord au niveau de la détection de la cible par le moteur de reconnaissance Spikenet, puis en raison du bruit et des erreurs de la centrale inertielle fournissant l'orientation des caméras (certains facteurs extérieurs peuvent par exemple perturber le magnétomètre). Cette erreur d'orientation est illustrée dans la Figure II-28 . Si α est l'angle correct par rapport au nord magnétique, alors une erreur angulaire β du capteur induit par conséquent une erreur (notée *e)* dans le calcul de la position de l'utilisateur, tel que représenté sur la Figure II-28-A. Celle-ci s'accroit de plus avec la distance *d* à l'objet, plus précisément, $e = 2d \times \sin(\beta/2)$. La Figure II-28-B illustre ce problème en simulant une erreur constante de la centrale inertielle de 40°. De ce biais résultent des positions de l'utilisateur incorrectes matérialisées en rouge, qui, comme on peut le voir, sortent de la trajectoire, contrairement aux vertes, ayant une orientation par rapport au nord correcte.

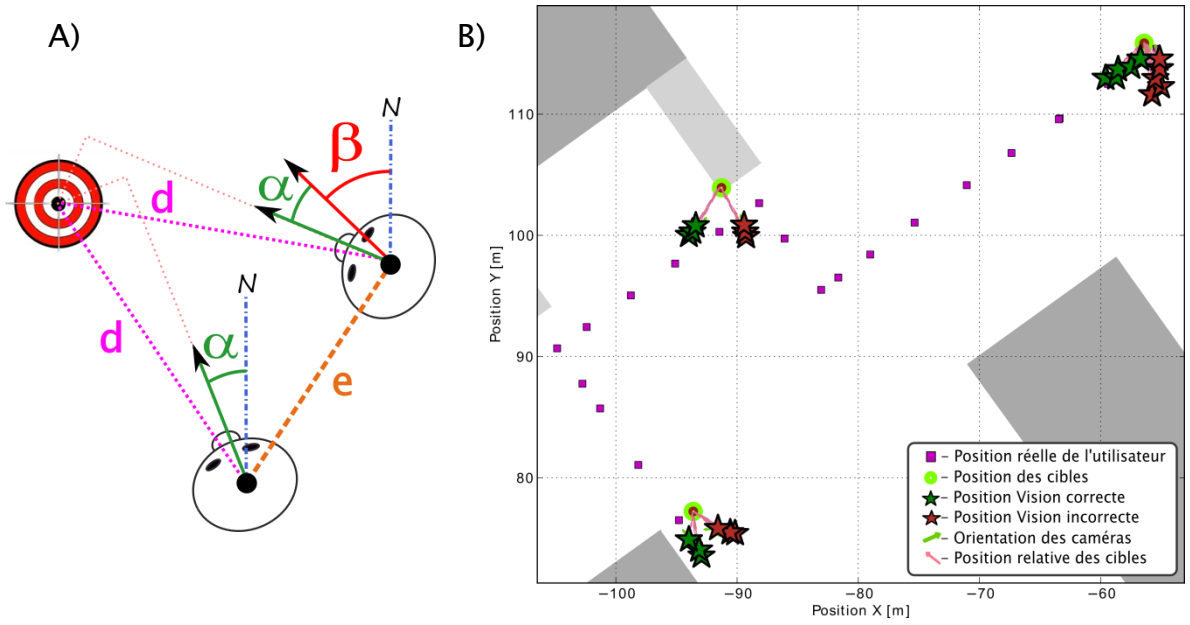


Figure II-28 Erreur de positionnement liée aux imprécisions du magnétomètre. A) L'erreur de position *e* est illustrée pour une erreur de cap β et une distance à la cible visuelle *d*. B) Simulation sur un parcours de test d'une erreur de cap constante (40°), résultant en un positionnement incorrect (en rouge).

Les derniers facteurs pouvant influer sur la précision de la localisation basée sur la vision sont liés à l'algorithme de stéréovision permettant le calcul de la distance de l'objet. Comme nous l'avons détaillé dans la section II.2.1.2, ce calcul se base sur la disparité d'un point entre l'image de gauche et l'image de droite. Pour cela il est d'abord nécessaire d'effectuer une mise en correspondance, c'est-à-dire retrouver le point correspondant à celui recherché dans la deuxième image. Cette recherche se base sur les caractéristiques visuelles du voisinage proche des points, et si ceux-ci sont uniformes ou peu caractéristiques il est

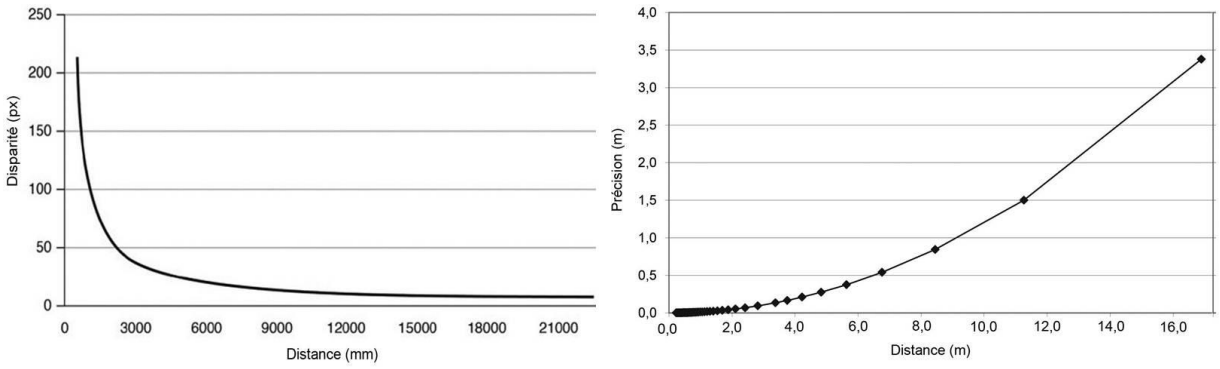


Figure II-27 Précision de la localisation par stéréovision en fonction de la distance (données constructeur).

probable que la localisation soit imprécise, entraînant par conséquent une erreur dans la valeur de disparité et donc de distance. De plus, même sous l'hypothèse d'une mise en correspondance absolument correcte, la précision décroit inévitablement avec la distance. En effet tel qu'illustré dans la Figure II-27, la disparité augmente exponentiellement lorsque la distance diminue. La précision d'estimation de la profondeur décroit par conséquent avec la distance, mais également avec la résolution de l'image, avec l'espacement des deux caméras, et avec la taille du champ de vision (plus la focale est faible plus la résolution diminue).

Afin d'estimer la fiabilité de ces mesures et leur utilisabilité dans notre contexte d'utilisation, nous avons procédé à une série de tests consistant à porter le dispositif en restant immobile en face d'une cible visuelle (4 différentes distances ont été testées : 2, 4, 8 et 10 mètres), puis à bouger la tête (donc les caméras, qui sont montées sur un casque) dans toutes les directions pendant soixante secondes. Parmi les différents essais, la cible a été reconnue au total plus de 1500 fois, soit environ 25 détections par seconde. La Figure II-29 montre que dans plus de 80% des cas, l'erreur de positionnement restait inférieure à 40 cm, une valeur qui semble donc adaptée au positionnement et au guidage d'un non voyant.

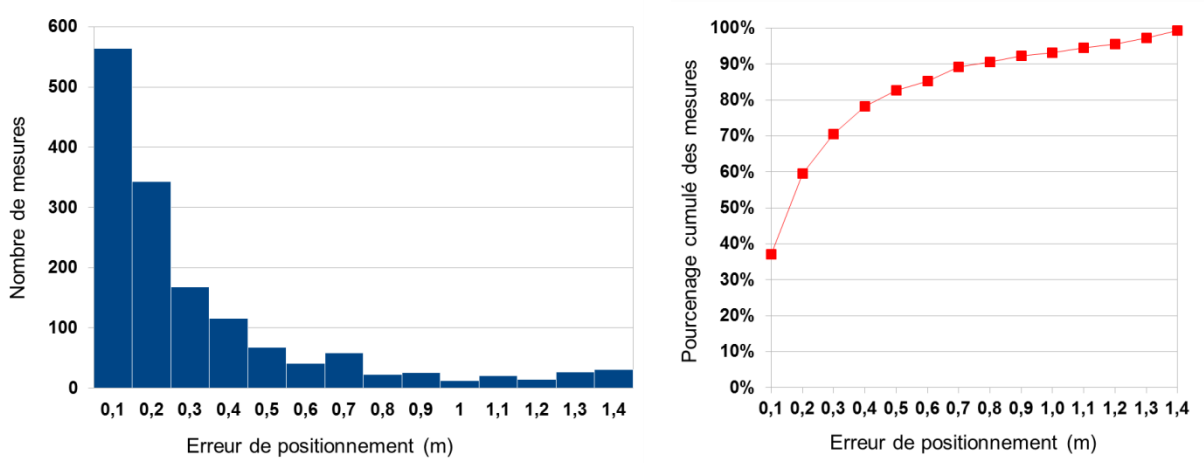


Figure II-29 Erreurs du positionnement basé sur la vision pour 1525 mesures effectuées à une distance de 2,4,8 ou 10 mètres d'un point visuel.

2.4.2 Position GPS

Nous avons utilisé au cours des différentes séries d'expérimentation deux dispositifs de navigation par satellite. Le premier, Angéo, est le fruit des travaux de l'entreprise Navocap. Il s'agit d'un boîtier comprenant un récepteur GPS, des capteurs inertiels, ainsi qu'un système embarqué effectuant les traitements de ces différents capteurs pour fournir une position consolidée, tirant parti des méthodes de navigation à l'estime. Les différents tests que nous avons effectués avec les prototypes mis à disposition par Navocap n'ont cependant pas été vraiment concluants. Le boîtier, toujours en cours de développement lors de ces tests, montrait en effet une erreur moyenne importante et des problèmes de lissage trop accentué de la trajectoire et de l'orientation, comme en témoignent les relevés de la Figure II-30.

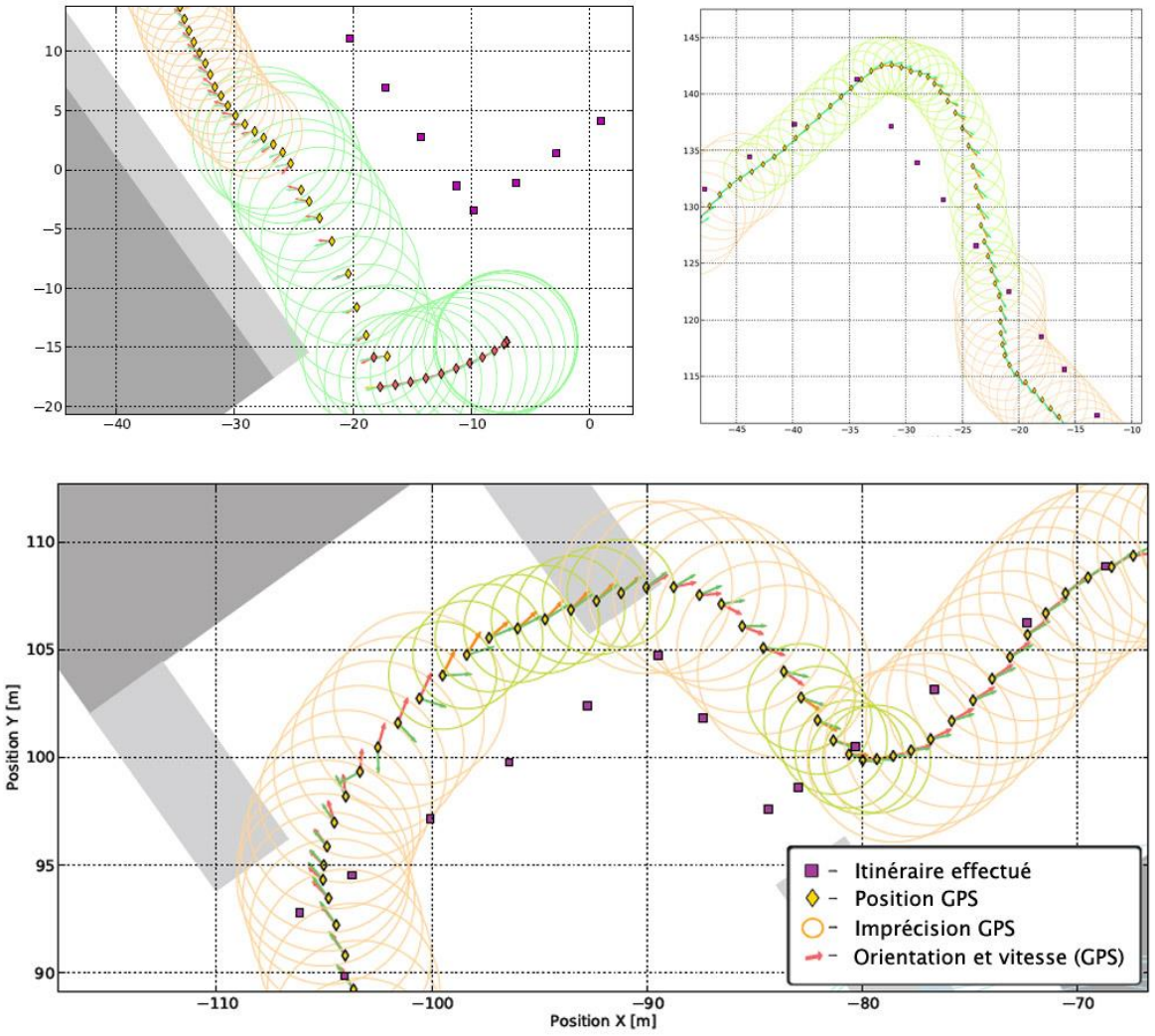


Figure II-30 Erreurs du GPS Angéo : (en haut à gauche) orientation incorrecte ; (en haut à droite) lissage de la trajectoire ; (en bas) lissage de l'orientation du corps.

Nous avons donc préféré nous tourner vers un récepteur GPS standard, en l'occurrence un récepteur de la marque Qstartz, de taille compacte, communiquant en Bluetooth avec le reste du système. A titre illustratif, deux exemples de traces GPS enregistrées avec ce récepteur sont présentées dans la Figure II-31, l'une assez précise, l'autre beaucoup moins[1], en particulier dans les zones abritées à faible couverture.

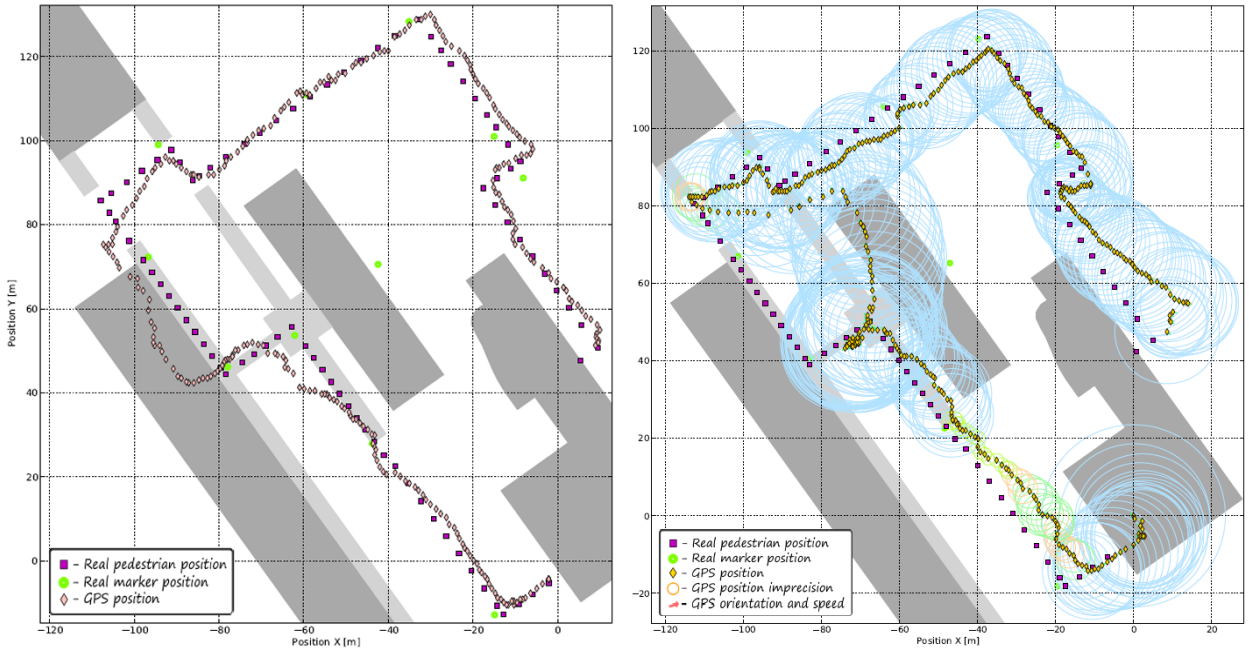


Figure II-31 Traces GPS relevées avec le GPS Qstartz.

Celui-ci n'intègre aucun algorithme particulier de correction de positionnement, mais fournit en plus de la position des informations sur la qualité du signal et donc de la localisation (plus précisément le nombre de satellites, la force du signal le plus fort ainsi que l'indice HDOP (Horizontal Dillution Of Precision), traduisant la précision horizontale en fonction de la répartition et de l'éloignement des satellites). Ces valeurs permettent de déterminer un indice de confiance sur le positionnement satellite afin de pouvoir pondérer les différentes sources d'information (vision, dead-reckoning, gps) selon la couverture satellite. Le dispositif fonctionnant à une fréquence d'1Hz, un module Ivy de relais (permettant l'interfaçage avec les autres agents) envoie donc toutes les secondes un message comme celui ci-dessous.

```
POS type=data temps=131006 lat=43.561584 long=1.467715 alt=140.2
vitesse=0.14816 cap=212.11 HDOP=0.79 Nb_Satellites=9 Force_Signal=46
```

[1] Ces différences entre les deux relevés peuvent s'expliquer par les conditions météorologiques ou la position des satellites lors de chacun des enregistrements.

Comme on peut le voir sur ce message, en plus de la position et des indices de qualité du signal, le récepteur GPS transmet également une valeur de cap et de vitesse, intégrées à partir des mesures GPS précédentes. Suite à nos premières observations sur ces données lors d'un comportement de marche, et aux fréquents changements d'orientations incorrects, nous avons décidé de mettre en place pour chacune un filtre de Kalman, interagissant l'un avec l'autre (la vitesse de déplacement dépendant de la vitesse angulaire et vice versa). La vitesse de déplacement ayant un caractère assez prédictible et lisse dans le cas d'un piéton, le filtrage permet ainsi de réduire le bruit de façon efficace. Pour l'orientation, nous avons fait l'hypothèse que lors d'un déplacement rapide, l'orientation du corps varie peu, alors qu'à l'inverse d'importants changements de direction peuvent intervenir lorsque la vitesse est nulle ou faible. La matrice Q du filtre de Kalman appliqué à l'orientation a donc été définie non comme une constante, mais comme fonction de la vitesse : $Q_{yaw} = 1/(speed + 1)^3$. En plus de ces filtrages, nous avons également ajouté des contraintes liées au modèle de marche, à savoir une vitesse maximale de 1,5 m/s et une accélération maximale de 0,7 m/s². Le résultat de ces méthodes est proposé dans la Figure II-32.

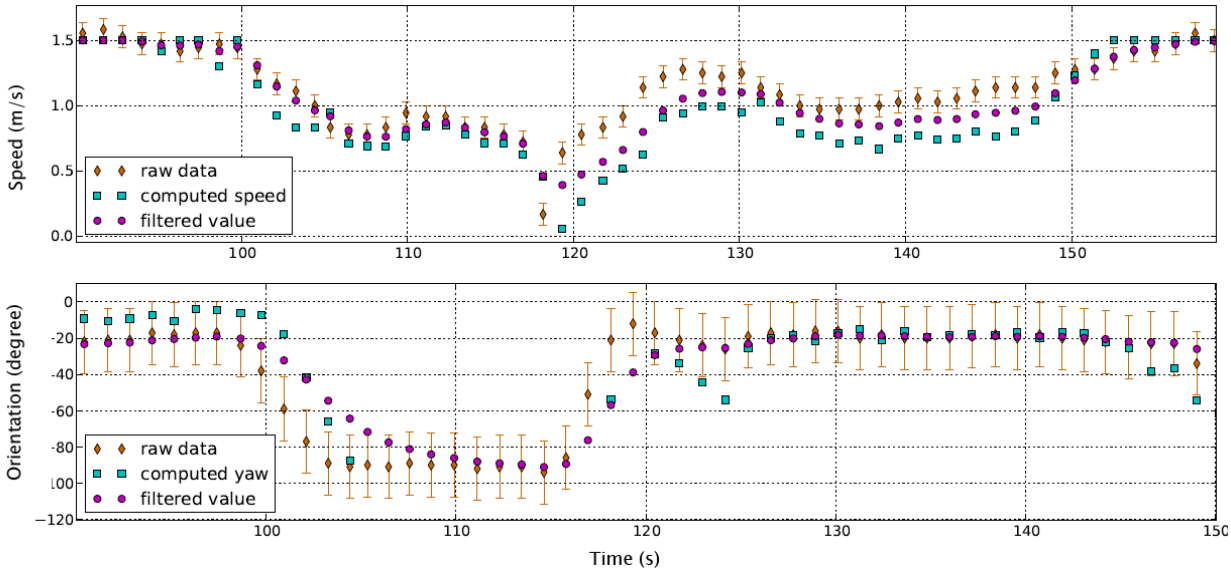


Figure II-32 Filtrage des données GPS. Durant un déplacement d'environ 1 m/s l'orientation filtrée suit les modifications du GPS. Lorsque le piéton ralentit, et donc que sa vitesse diminue, la fenêtre temporelle du filtre réduit en fonction. Enfin à l'inverse, lorsque la vitesse est au-delà d'1 m/s, l'orientation filtrée s'approche d'une valeur fixe.

Enfin, comme ces mesures de cap et de vitesse sont calculées à partir des données GPS, elles seront nécessairement dépendantes de la couverture satellite. Nous avons donc exploité les informations concernant la force du signal afin de pondérer et corriger celles-ci. Une mesure de confiance a donc été définie telle que :

$$confidence = \frac{1}{1 + e^{-0.25 \times (signal - 32)}}$$

A partir de celle-ci, nous pouvons pondérer la vitesse fournie par le GPS avec la vitesse de marche moyenne d'un piéton ($speed_{mean} = 0.9\ m/s$) de sorte que :

$$speed = (confidence \times speed_{GPS}) + \big((1 - confidence) \times speed_{mean}\big)$$

Et nous avons également deux indices d'imprécision de l'orientation et de la position qui seront utilisés dans le processus de fusion final :

$$position_{imprec} = (1 - confidence) \times 50$$

$$orientation_{imprec} = (1 - confidence) \times 35$$

2.4.3 Prétraitements

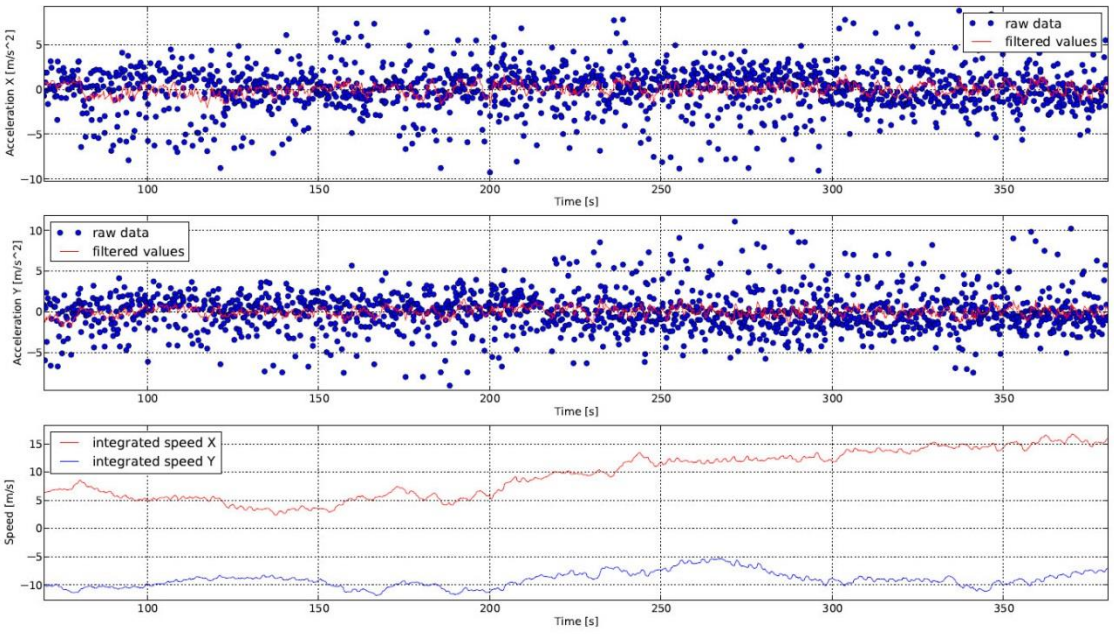


Figure II-33 Vitesse intégrée (en bas) à partir des valeurs filtrées d'accélération (en haut) du capteur Xsens positionné sur la hanche de l'utilisateur.

La phase de prétraitements consiste en un premier filtrage, relativement basique des différentes entrées du système afin d'en réduire le bruit, d'apporter certaines premières corrections et contraintes aux valeurs brutes reçues. En plus des filtres appliqués aux données du GPS mentionnées précédemment nous avons également appliqué un filtrage de Kalman au capteur inertiel positionné sur la tête, ainsi qu'au cap donné par celui positionné sur la hanche, celui-ci étant paramétré en fonction de la vitesse de marche fournie par le GPS. Un dernier filtrage a été utilisé pour les données d'accélération du capteur sur la hanche, afin d'estimer la vitesse angulaire utilisée ensuite pour l'estimation de la marche. Celles-ci sont présentées dans la Figure II-33.

2.4.4 Autocorrection

Les mécanismes d'autocorrection que nous avons intégrés à l'algorithme de fusion ont été inspirés des techniques de SLAM [Bailey and Durrant-Whyte, 2006; Durrant-Whyte and Bailey, 2006], consistant en l'interaction de processus de localisation et de cartographie. Les méthodes de SLAM visent donc à simultanément créer une carte de l'environnement à partir de différents capteurs (souvent visuels ou laser) et à estimer sa position dans celle-ci. Elles sont l'objet de nombreuses recherches et application dans le milieu de la robotique depuis près d'une vingtaine années, et reposent sur l'extraction d'invariants de l'environnement et l'estimation de son mouvement au sein de celui-ci.

Dans l'architecture du moteur de fusion, les différents sous modules intermédiaires d'autocorrection interagissent ainsi entre eux dans des boucles similaires. A partir des donnés prétraitées de la phase précédente et de ces interactions mutuelles, l'étape d'autocorrection vise à fournir une estimation fiable de l'orientation et de la vitesse de l'utilisateur, ainsi que deux mesures indépendantes de sa position, l'une basée sur la vision, l'autre sur le GPS. Certains modules clés de cette étape seront détaillés ici. Pour plus d'informations se reporter à [Borovec, 2011; Borovec et al., 2014; Brilhault et al., 2011].

Décalage GPS

Lors de nombreux relevés GPS nous nous sommes aperçus que la trajectoire mesurée correspondait assez précisément à la trajectoire réelle, à l'exception d'une translation (ou parfois d'un facteur d'échelle). Lorsqu'un point visuel est détecté, et que l'on obtient une position fiable basée sur la vision, un vecteur de correction est donc calculé entre celle-ci et la dernière position GPS reçue, qui sera ensuite appliqué à toutes les positions GPS ultérieures, jusqu'à la prochaine reconnaissance d'un point visuel. Ce procédé introduit certes quelques discontinuités dans la trajectoire, mais permet un gain très important sur la précision de localisation comme illustré dans la Figure II-34.

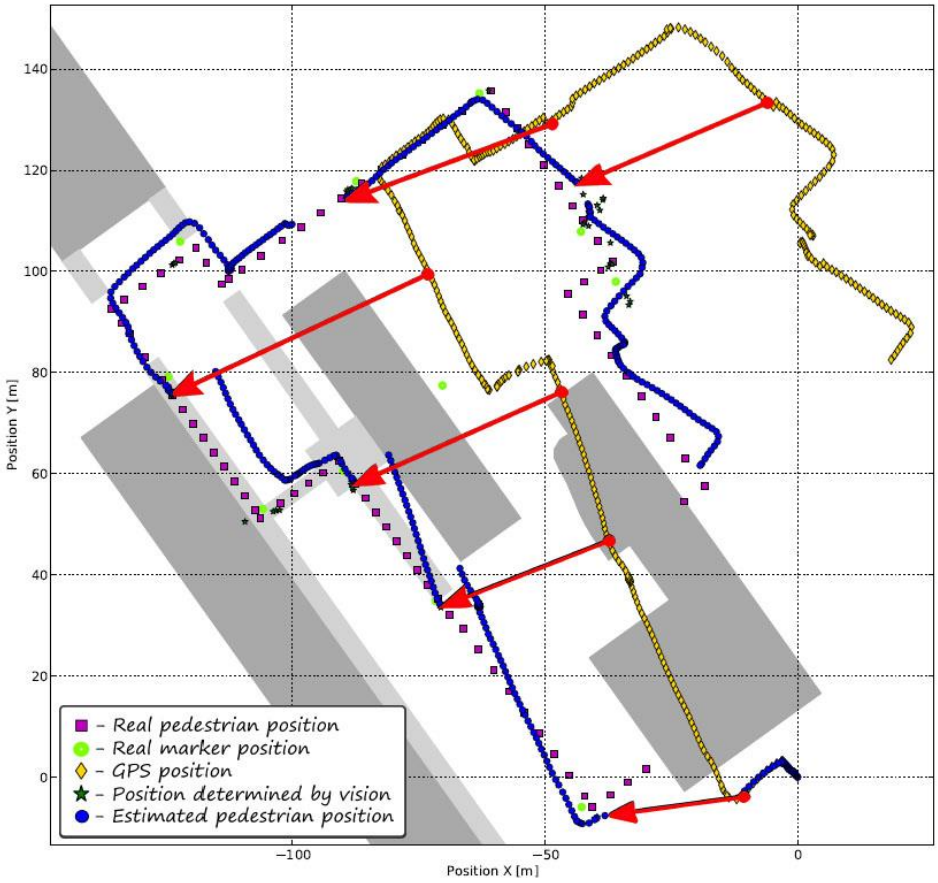


Figure II-34 Correction du GPS. La trajectoire en jaune correspond au GPS seul, celle en bleu au GPS utilisant le recalage basé sur la vision (les flèches rouges correspondent à chacun de ces recalages, intervenants lors de détections visuelles)

Raffinement de l'orientation du corps

L'estimation du cap fournie par les magnétomètres de la centrale inertielle Xsens placée à la hanche est, comme tous les compas, soumise à l'influence magnétique des éléments environnant. Pour tenter de corriger ces erreurs ce composant intègre les informations provenant du GPS ainsi que des positions reconstruites grâce à la détection de points visuels. Deux mécanismes de corrections interviennent. Le premier utilise les informations de cap fournies par le GPS pour ajouter une constante de recalage aux valeurs du Xsens. Comme nous l'avons expliqué, ce cap GPS n'est fiable que lorsque la précision GPS est haute, les corrections ne sont donc appliquées que lorsqu'une séquence suffisante de mesures consécutives du GPS possède une confiance au-dessus d'un certain seuil (nous avons expérimentalement fixé ces valeurs à 11 relevés dont la confiance est supérieure à 85 %). Lorsque c'est le cas, la moyenne des caps GPS fournies lors de cette séquence et celle des orientations Xsens correspondantes sont calculées pour déterminer la valeur de correction à appliquer.

La deuxième méthode repose sur la comparaison de deux vecteurs, l'un correspondant au mouvement relatif entre les positions calculées à partir de deux détections successives de points visuels, l'autre à celui estimé à partir de la centrale inertielle et du GPS. La différence angulaire entre ceux-ci permet de déterminer la correction du cap Xsens, et la différence de ratio la correction de la vitesse. La Figure II-35 montre clairement l'intérêt de cette méthode, où le trajet corrigé correspond presque parfaitement à celui effectué.

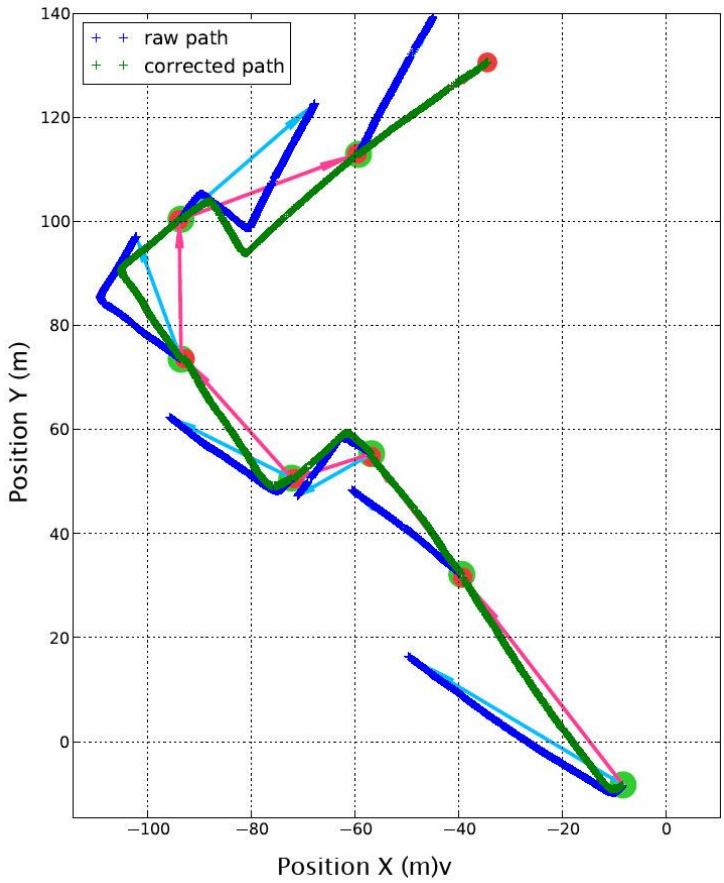


Figure II-35 Correction de l'orientation. La comparaison des vecteurs roses entre les positions calculées grâce aux détections visuelles et des vecteurs bleus clair calculés à partir des centrales inertielles et du GPS permet la correction de l'orientation et de la vitesse, et donc un trajet rectifié symbolisé en vert.

Estimation du mouvement

A partir de ces différentes corrections apportées au cap et à la vitesse issues conjointement des informations GPS, visuelles et inertielles, ce module permet l'estimation finale du vecteur de déplacement de l'utilisateur. Il prend également en compte l'état de marche (à l'arrêt, ou en mouvement), calculé à partir de l'enveloppe des signaux

d'accélération de la centrale inertielle à la hanche, et comporte un dernier filtre de Kalman permettant de lisser les données en sortie. La figure présente les résultats de ces différentes méthodes de correction.

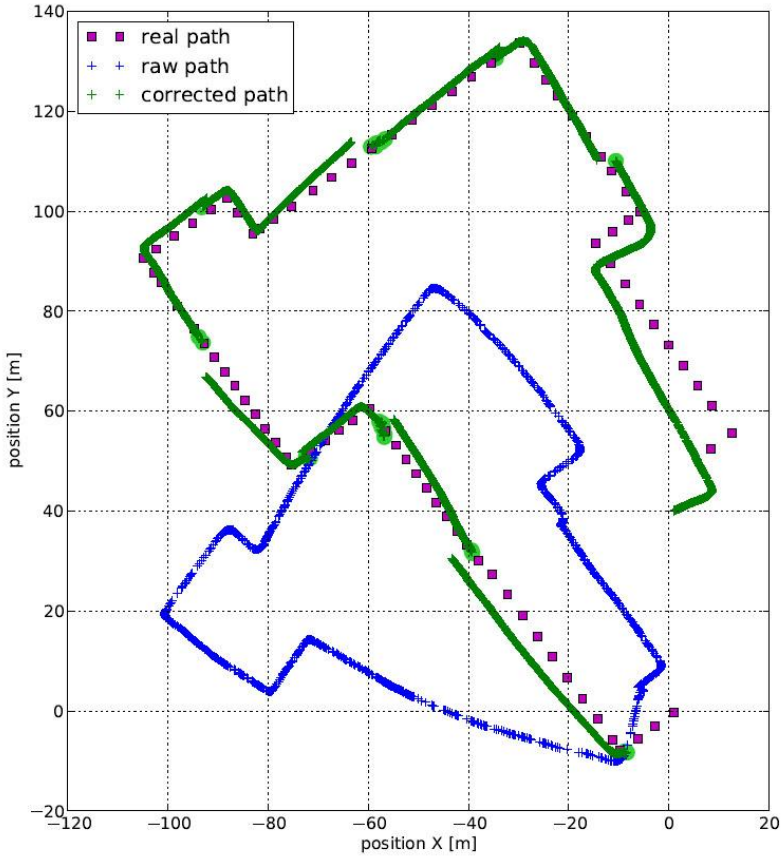


Figure II-36 Mouvement relatif calculé à partir de la vitesse et du cap (brutes en bleu, et corrigées en vert) avec recalage par les détections de Points Visuels.

Fiabilité des détections visuelles (RecoPath)

Comme nous l'avons expliqué dans la section II.2.3.3, le chargement des modèles correspondants aux points visuels à rechercher dans l'environnement est conditionné par la dernière position connue de l'utilisateur. Il s'agit généralement d'un simple rayon (par exemple tous les PVs à moins de 50 mètres), et pour certains amers d'une valeur déterminée par les propriétés de la cible (des bâtiments de taille importante pouvant être reconnus à une distance bien plus grande que celle par défaut, ou à l'inverse d'autres objets plus petits ne pouvant être détectés qu'à quelques mètres). Ces contraintes, en plus de diminuer le temps de traitement, permettent également d'éviter de nombreuses fausses détections. Il arrive néanmoins que, malgré ces mécanismes, certaines fausses alarmes interviennent occasionnellement. Comme au sein du moteur de fusion de nombreuses corrections sont

conditionnées par ces détections visuelles, de fausses alarmes, même rares, peuvent avoir un impact important. Pour améliorer la localisation, nous avons donc cherché une méthode permettant de filtrer ces fausses détections, et la solution retenue, inspirée du filtrage particulaire, a été baptisée RecoPath.

L'algorithme se base sur la cohérence d'une séquence de détections, et sur le

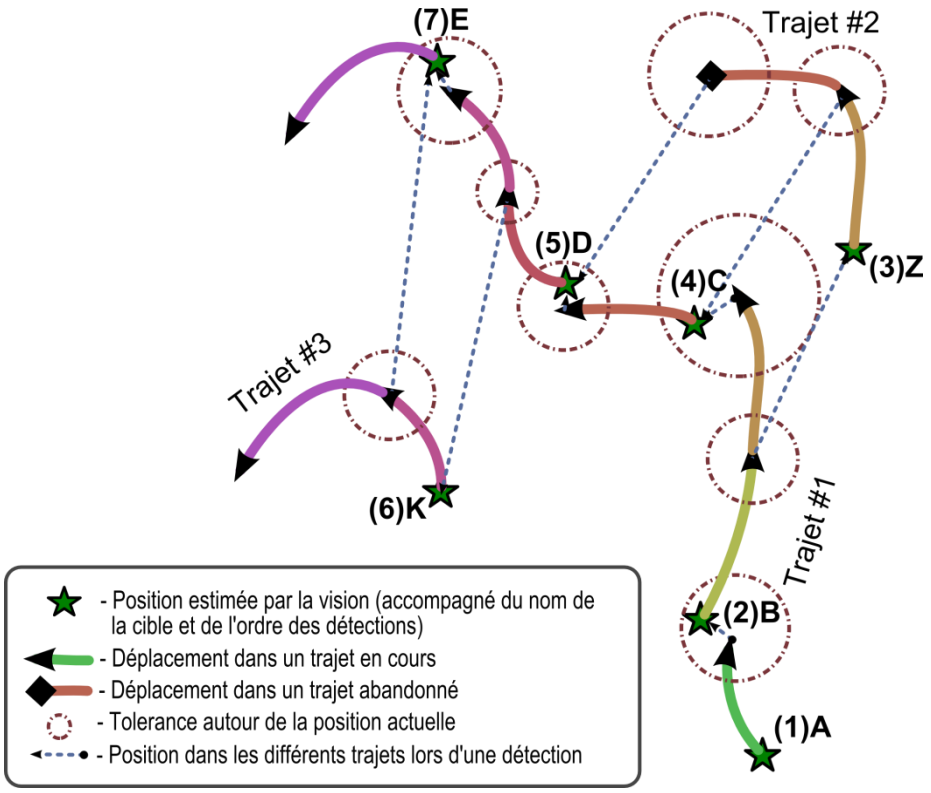


Figure II-37 Illustration de la méthode RecoPath. Différents trajets peuvent être créés ou terminés selon leur probabilité. Ceux-ci sont calculés par les détections visuelles et les déplacements entre elles. Dans cet exemple le premier trajet débute à la détection du PV A, lorsque le point B est reconnu, la position reste dans la tolérance autorisée, et il est donc ajouté à ce trajet. En revanche à la détection du point Z, en dehors de la portée courante, un nouveau trajet est créé (bien que le trajet 1 reste actif, ayant la probabilité la plus haute). Les détections des PVs C et D sont ensuite ajoutées au premier trajet, et, lors de cette dernière le trajet 2, trop peu probable, est alors terminé. La détection de la cible K entraîne la création d'un nouveau trajet, pour les mêmes raisons que le trajet 2 précédemment. Et finalement la dernière cible, étant de nouveau cohérente avec le trajet 1 est ajoutée à celui-ci, renforçant ainsi encore sa probabilité.

mouvement relatif entres celles-ci, estimé grâce à l'orientation et à la vitesse de l'utilisateur (une fois appliquées les corrections détaillées précédemment). Ce procédé n'est au final pas si éloigné du comportement de navigation d'un voyant. Imaginons que l'on suive un trajet

dans un environnement inconnu à l'aide d'une carte. Si, alors qu'on pense être dans la bonne direction, on aperçoit un objet, ou un point de repère n'étant pas supposé être sur l'itinéraire, il est probable que l'on n'y prête pas attention et que l'on continue sur la même route. En revanche, après avoir constaté la présence de plusieurs éléments incohérents, on reconsidèrera probablement sa position pour se réorienter en fonction de ceux-ci. La méthode RecoPath fonctionne sur le même principe. Une liste des différents trajets possibles est mise à jour en fonction des détections et des mouvements de l'utilisateur. Ainsi lors d'une nouvelle détection visuelle, celle-ci est soit ajoutée à un trajet existant (si la position reconstruite est suffisamment proche de la position de la dernière cible de ce même trajet, à laquelle on a appliqué le déplacement effectué depuis), soit ajoutée à un nouveau trajet dans le cas contraire. A chaque trajet est associé une probabilité, également mise à jour à chaque nouvelle détection. Celle-ci dépend principalement de la longueur de la séquence de détection. Au cours de ces mises à jour, l'algorithme pourra changer de trajet « actif » (c'est-à-dire le trajet le plus probable), et supprimer les trajets dont la probabilité serait devenue trop faible.

Des expériences préliminaires nous ont permis de fixer les équations et paramètres de ces probabilités associées à chaque trajet. La corrélation entre la longueur des séquences de cibles consécutives cohérentes et la confiance en celles-ci semblant suivre une loi Gamma, nous avons donc modélisé la probabilité des trajets à partir de trois éléments utilisant la fonction de répartition G_{cdf} (*Gamma Cumulative Distribution Function*) donnée ci-dessous :

$$G_{cdf}(x;k,\theta) = 1 - \sum_{i=0}^{k-1} \frac{1}{i!} \cdot \left(\frac{x}{\theta}\right)^{i} \cdot \exp\left(-\frac{x}{\theta}\right)$$

- La probabilité P_{active} d'une séquence de points visuels d'être à l'état actif (lorsque la dernière détection vient d'être ajoutée à ce trajet) ou à l'état inactif (P_{other}) :

$$P_{active} = 0.5 + G_{cdf}(n, 4{,}2)/2$$

$$P_{other} = 0.5 + G_{cdf}(n, 2{,}9)/2$$

- Le coefficient de pondération dépendant de la longueur n de la séquence ininterrompue de détections : $P_{long} = 0.5 + G_{cdf}(n, 1{,}6)/2$
- La probabilité antérieure de la séquence P_{group}, de sorte que la nouvelle valeur P (utilisant P_{active} ou P_{other} selon s'il s'agit du trajet auquel a été ajouté la détection ou d'un autre) soit :

$$P = P_{active/other} \times P_{long} \times P_{group}$$

Sur un parcours test (le trajet sur le campus de l'université présenté dans la section résultats), l'efficacité du RecoPath à correctement séparer les fausses détections des valides a été mesurée. La totalité des fausses détections a bien été exclue par l'algorithme, et parmi les vrais positifs 23,7% ont été incorrectement rejetés. La perte de certains points visuels considérés comme fausses alarmes a néanmoins un impact beaucoup plus faible que les fausses détections, et cette méthode s'avère donc particulièrement utile, telle qu'en témoigne la Figure II-38, où l'on peut observer les erreurs de positionnement qu'auraient induit ces détections écartées par le RecoPath.

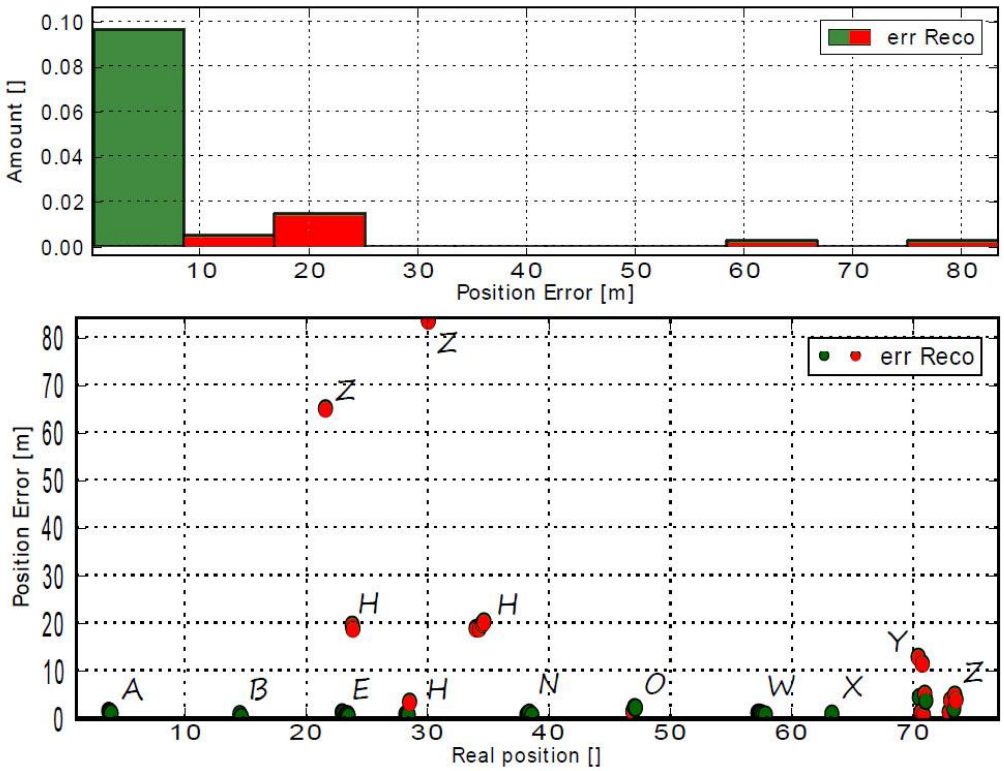


Figure II-38 Application du RecoPath (en rouges sont indiqués les détections classées comme incorrectes, en vert les retenues).

Positionnement final

Pour l'intégration de plusieurs sources de données en vue d'une fusion et de l'estimation du positionnement, les deux méthodes les plus couramment utilisées sont les filtres de Kalman et les filtres particulaires, dont la mise en place pour la localisation d'un piéton est par exemple proposée dans [Ceranka, 2002; Ceranka and Niedzwiecki, 2003]. Il existe une grande variété de filtres dérivant de ces des familles : *Extended Kalman Filter, Unscented Kalman Filter, Hybrid Kalman Filter, Sequential Importance Sampling Particle filter, Local Likelihood Sampling Particle filter, Rao-Blackwellised Particle filter, Gaussian Particle*

filter,... Nous n'aborderons pas ici les détails de chacun et leurs différences, pour cela se reporter à [Kaplan and Hegarty, 2006; Ristic et al., 2004].

Nous avons donc décidé d'implémenter ces deux méthodes, plus précisément un filtre particulaire avec rééchantillonnage par importance (SIP PF), et un filtre de Kalman hybride. Ils utilisent en entrée la position estimée par la vision, celle par le GPS (avec un indice de confiance associé), ainsi que le cap et la vitesse. Ces différentes valeurs ayant déjà fait l'objet de filtrages et de corrections, la phase de fusion finale est relativement classique. Les détails de l'implémentation ainsi que le paramètres utilisés, déterminés expérimentalement, sont disponibles dans [Borovec, 2011; Borovec et al., 2014].

La comparaison des deux méthodes, proposée dans les Figure II-39 et Figure II-40, donne l'avantage au filtre de Kalman, bien que les résultats soient très proches. Cependant ces algorithmes sont extrêmement variables selon les modèles et les paramètres utilisés. Si nous avons finalement retenu le filtre de Kalman pour ses performances et sa rapidité d'exécution, il n'est pas impossible que le filtre à particules ait pu donner des résultats semblables voir supérieurs en ajustant et en optimisant ses paramètres.

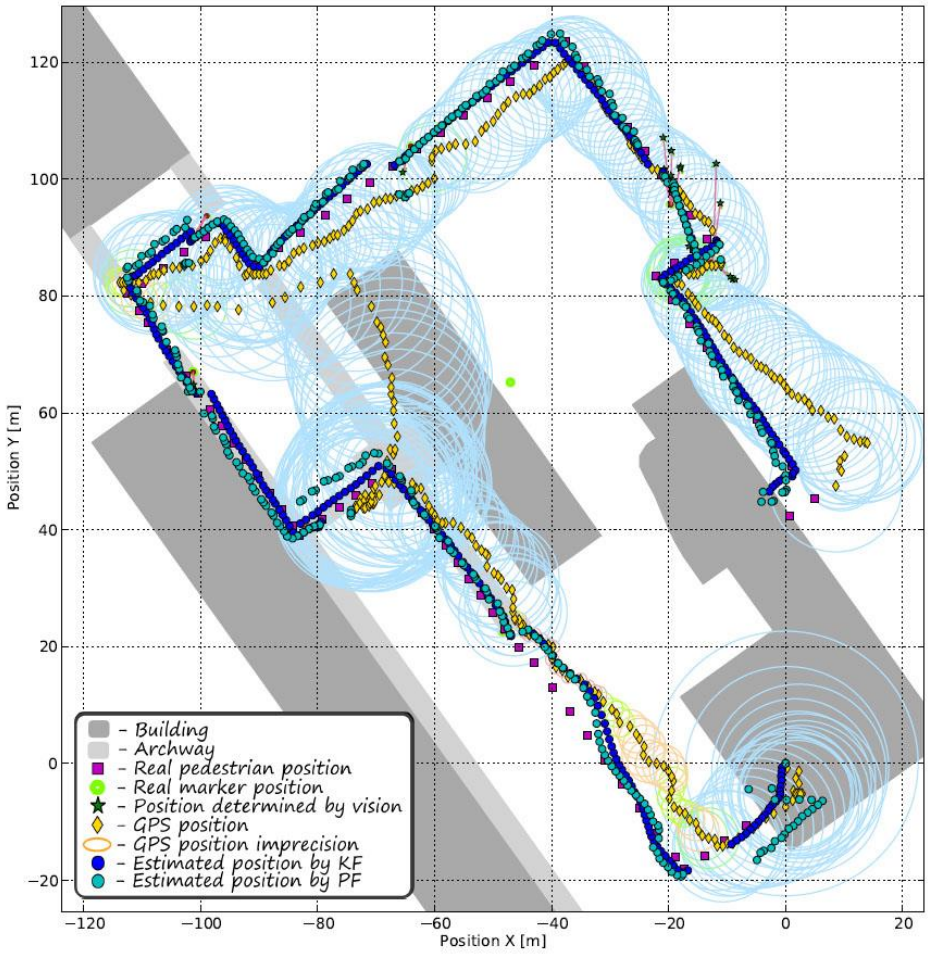


Figure II-39 Comparaison des trajectoires estimées par le filtre à particules (en turquoise) et le filtre de Kalman (en bleu)

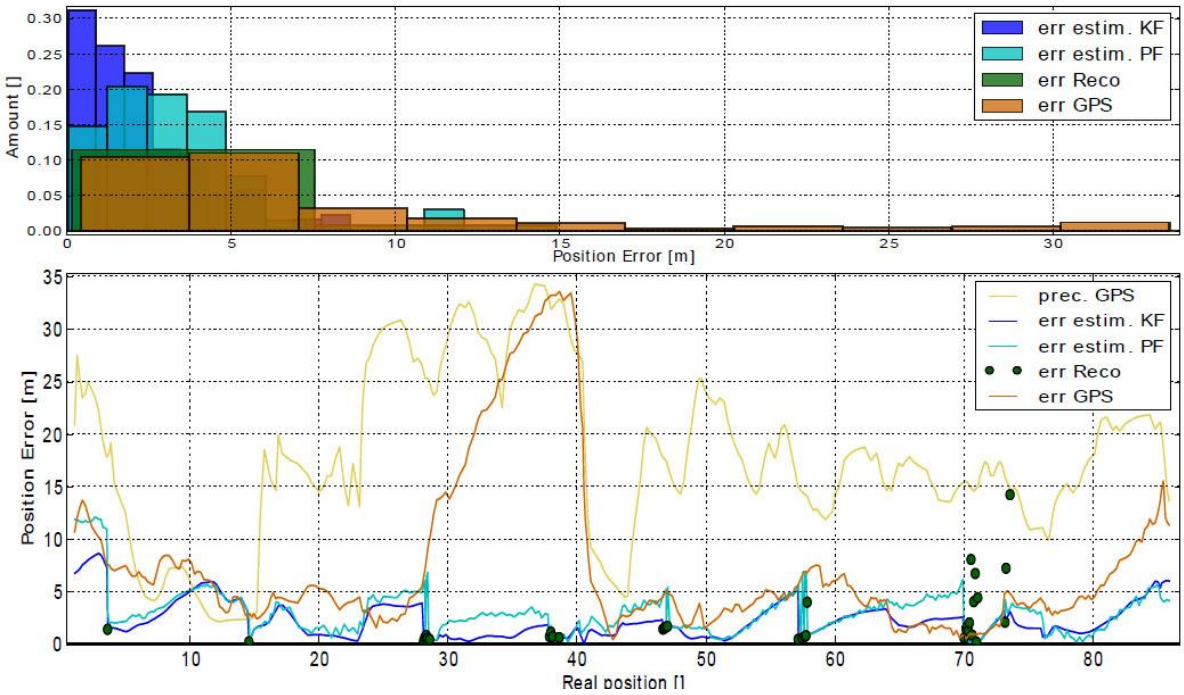


Figure II-40 Comparaison de l'erreur du filtre à particules (en turquoise) et du filtre de Kalman (en bleu)

2.5 Résultats

Pour mettre en place puis valider le module de fusion, de nombreux tests ont été pratiqués dans différents environnements. Deux trajets ont finalement été retenus pour l'expérimentation du système, le premier au sein du campus de l'université, et le second dans le centre-ville de Toulouse. Ils constituent des parcours réalistes de locomotion pour un piéton, comportant des traversées de route, des sections de longueur variable, divers changements de direction, et ont été choisis de façon à couvrir un maximum de situations pouvant être rencontrées lors de l'utilisation du dispositif en conditions réelles, à savoir des aires dégagées, des canyons urbains, ou encore des zones sans couverture satellites.

De nombreux parcours ont été réalisés pour chacun des trajets, durant lesquels l'ensemble des données recueillies ont été collectées afin de pouvoir être analysées et rejouées en variant les paramètres de l'algorithme de fusion. Ces nombreux jeux de donnés, obtenus dans des conditions changeantes, ont aussi permis d'observer les variations induites par les conditions atmosphériques sur le GPS, ainsi que par les changements de luminosité au niveau de la reconnaissance visuelle des cibles (selon l'heure de la journée, les ombres et l'éclairage peuvent en effet modifier l'apparence de certaines éléments de l'environnement de façon sensible). Afin de pouvoir valider l'algorithme de fusion il était nécessaire d'avoir comme référence la position réelle de l'utilisateur. La trajectoire seule

n'est pour cela pas suffisante. Lorsque la position estimée s'éloigne de celle- ci, la mesure de l'erreur doit en effet se baser sur sa position réelle à ce moment précis, et non sur le point de la trajectoire le plus proche. Nous avons donc marqué au sol des points, espacés de 5 à 10 mètres généralement, dont les coordonnées exactes ont été saisies dans le SIG. Enfin, lors des collectes de donnée, l'expérimentateur devait presser un bouton lors du passage sur chacun de ces points. Pour estimer la position réelle tout au long du trajet nous avons fait l'hypothèse d'une vitesse constante entre ces points, un choix offrant une précision suffisante étant donné leur proximité.

2.5.1 Expérimentation campus

Le premier parcours, d'une longueur de 360 mètres, a été réalisé sur le campus de l'université Paul Sabatier. Il comporte 10 changements de direction, deux traversées de route, des sections dégagées où la précision attendue du GPS devrait être maximale, ainsi que des passages sous des préaux masquant à l'inverse la couverture satellite. Les points marqués au sol (86 au total) et validés manuellement au cours du trajet pour obtenir la position réelle de l'utilisateur ont été disposés environ tous les 5 mètres (voir Figure II-41).

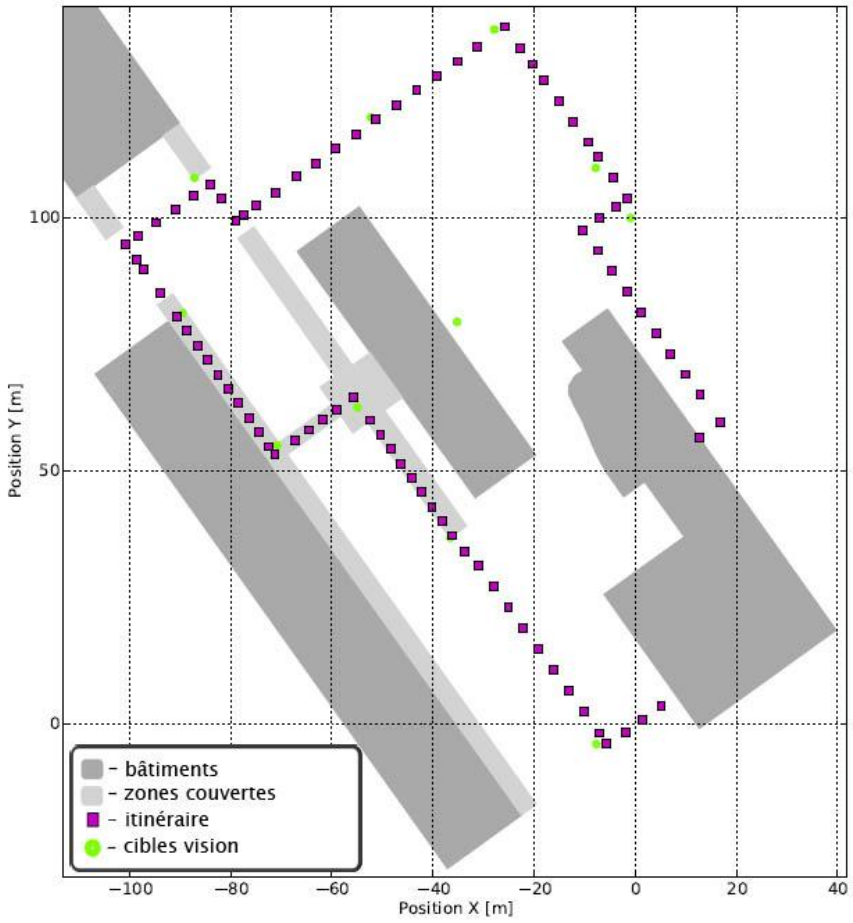


Figure II-41 Itinéraire du parcours Campus.

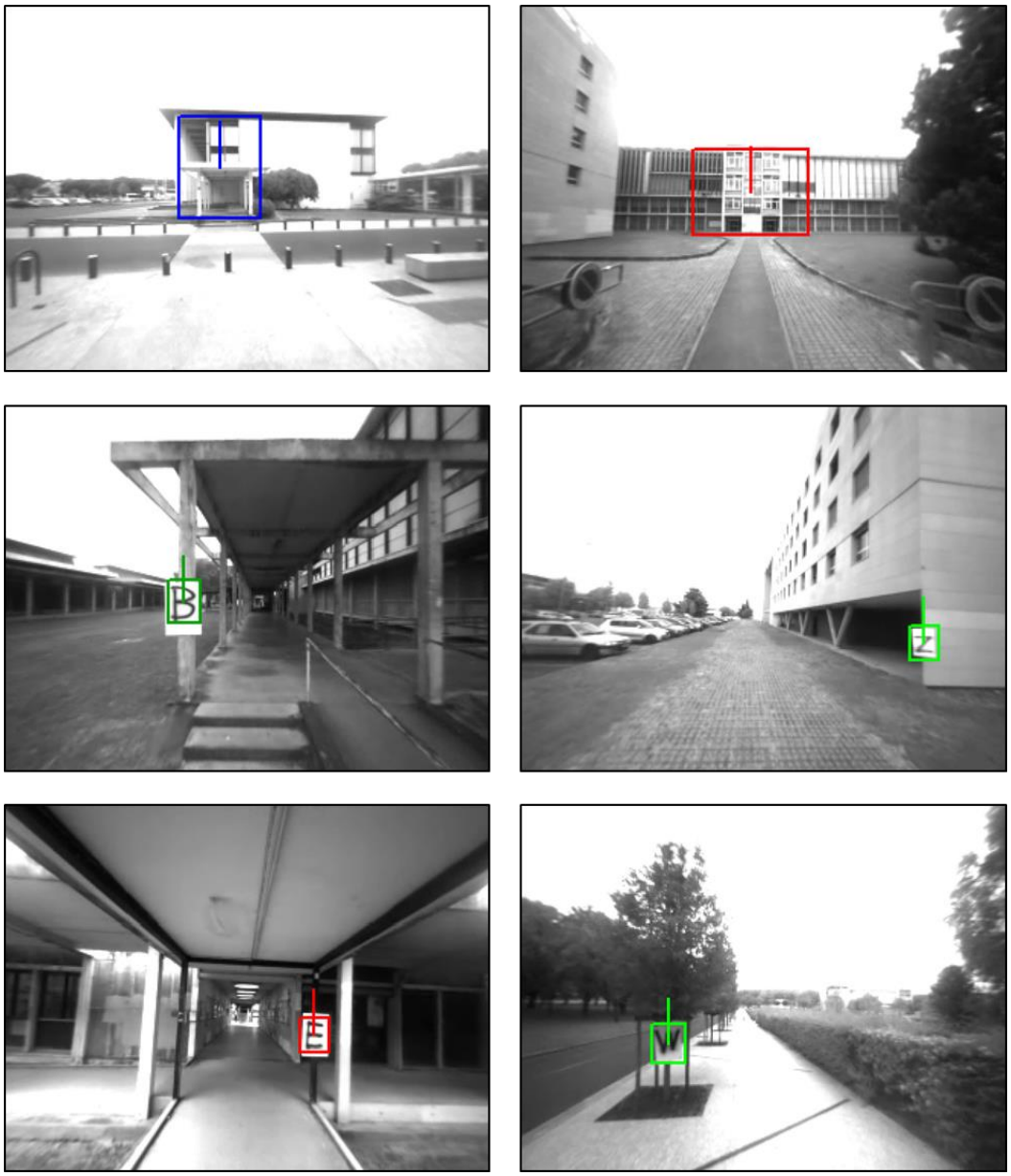


Figure II-42 Exemples de cibles visuelles du parcours sur le campus

La sélection de cibles visuelles dans l'environnement permettant la correction du positionnement a été délicate pour ce premier trajet. La plupart des espaces autour du parcours étant assez ouverts et vides, il était difficile de trouver des éléments pouvant remplir le rôle de points visuels. De plus les quelques bâtiments environnants ont la plupart du temps des motifs très répétitifs, tout comme les préaux traversés. Par conséquent, si quelques cibles naturelles[1] ont pu être apprises (principalement des façades de bâtiments avec une configuration spatiale reconnaissable), nous avons fait le choix d'ajouter un certain

[1] C'est-à-dire existantes dans l'environnement.

nombre de cibles artificielles afin de pouvoir tester et valider l'algorithme de fusion, tout en restant sur ce site, commode par son emplacement (autour du laboratoire, ne nécessitant donc pas de se déplacer en centre-ville). Une dizaine de pancartes de 80 par 60 cm ont été positionnées au cours du trajet (tous les 30 mètres environ). Pour qu'elles puissent être apprises par l'algorithme de vision il fallait que chacune comporte un motif différent et suffisamment caractéristique. Nous avons décidé d'utiliser des lettres de l'alphabet, mais ce choix est complétement arbitraire, tout autre motif complexe aurait pu convenir. Quelques exemples des cibles utilisées sont présentés dans la Figure II-42.

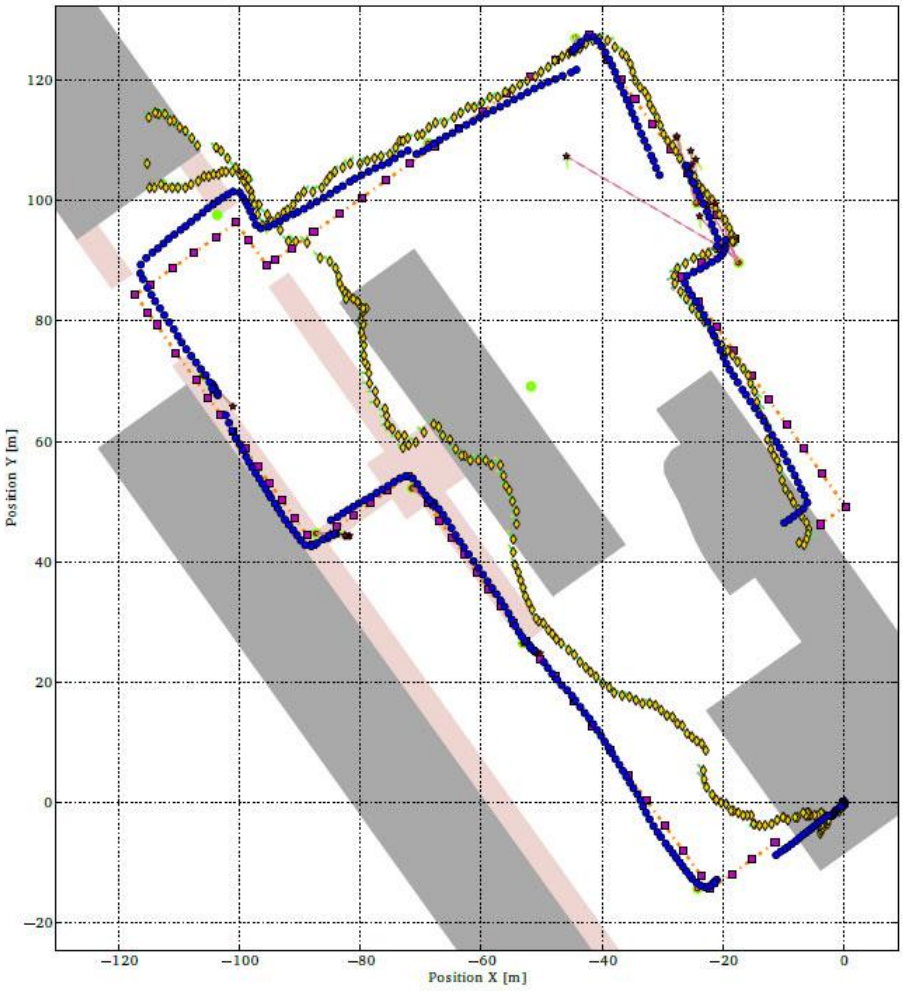


Figure II-43 Relevé du parcours Campus. Les points en violet correspondent au trajet effectué, ceux en jaune aux positions GPS, et ceux en bleu aux positions résultant de la fusion.

Les résultats de la trajectoire obtenue sont présentés dans la Figure II-43. Comme on peut le voir, le positionnement GPS brut, s'il est assez précis sur la fin du trajet où l'environnement est plutôt dégagé, présente une erreur significative en début de parcours, puis des performances très mauvaises dès le passage sous le préau, du fait de la

dégradation des signaux satellites. En moyenne, si l'imprécision reste inférieure à 15 mètres environ 75% du temps, elle peut néanmoins atteindre à certains moments 35 mètres (5% des mesures avaient entre 27 et 35 mètres d'erreur). Le positionnement fusionné, illustré en bleu, reste en revanche très proche de la trajectoire réelle tout au long du parcours et démontre l'intérêt de l'utilisation des amers visuels et des différentes corrections détaillées précédemment. Cette position corrigée reste en dessous de 2 mètres d'erreur plus de 50% du trajet, entre 2 et 6 mètres pour 40%, et seulement 10% des relevés présentent une erreur de 6 à 10 mètres, tel que visible dans la Figure II-44.

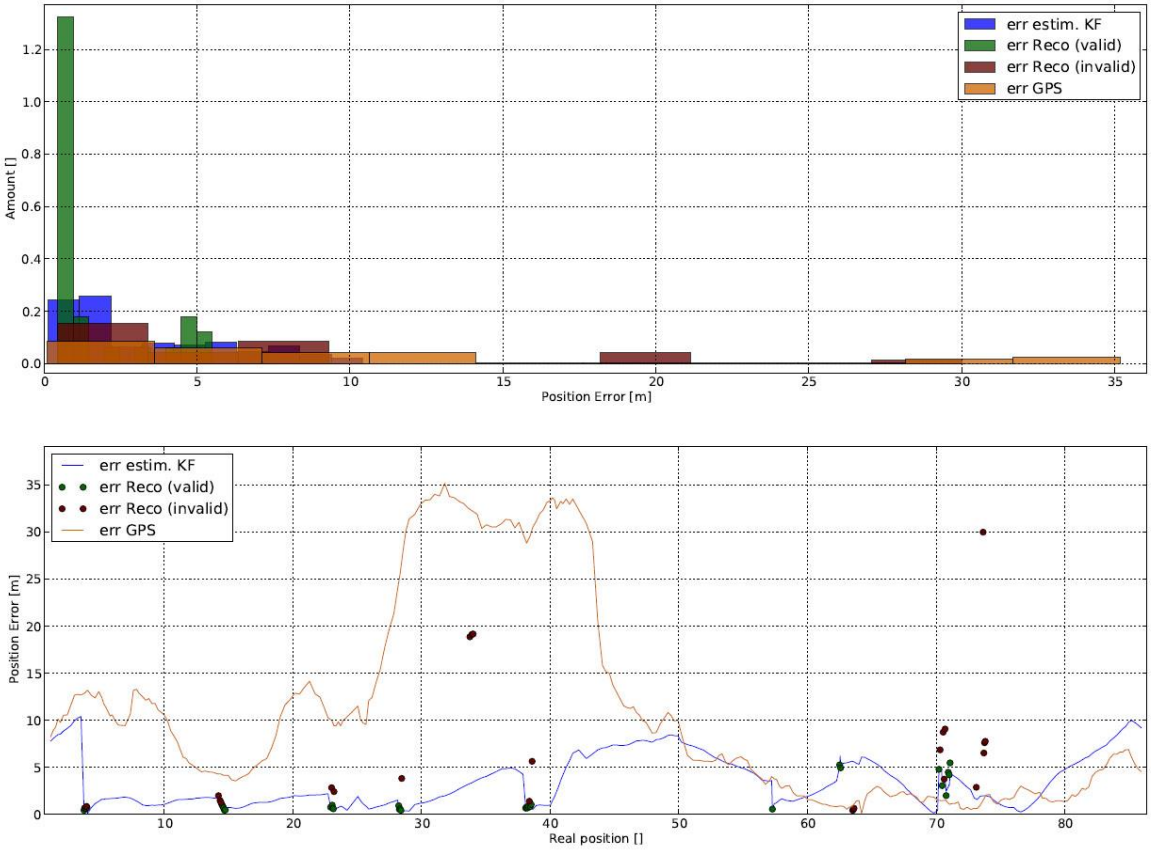


Figure II-44 Distribution des erreurs de positionnement au cours du trajet sur le Campus (erreurs du GPS, des détections visuelles correctes, incorrectes, ainsi que de la position fusionnée finale estimée avec un filtre de Kalman).

2.5.2 Expérimentation centre-ville

Ce second parcours, a été réalisé dans le centre-ville de Toulouse, au sein du quartier des Carmes. Son tracé est illustré dans la Figure II-45. Cet environnement, à l'inverse de l'Université, comprend une forte densité de bâtiments relativement hauts (au-delà de 15 mètres) et proches les uns des autres, des rues étroites, et peu d'aires dégagées. L'itinéraire de 660 mètres comprend 7 changements de direction, et est constitué de 75 points de validation espacés d'environ neuf mètres.

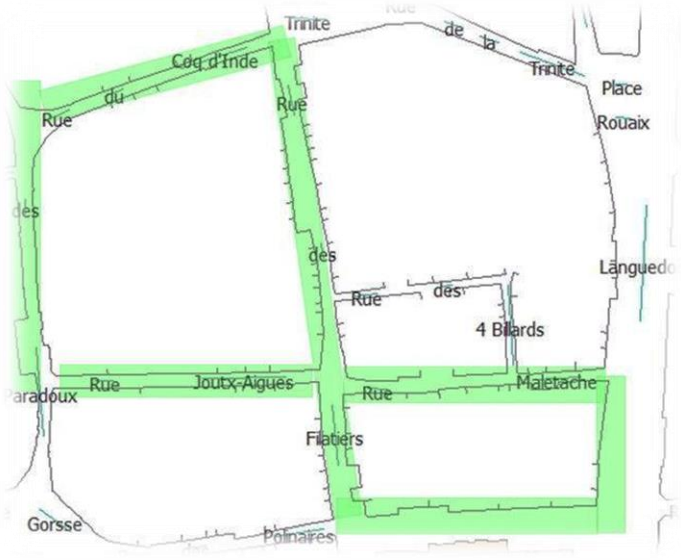


Figure II-45 Plan du parcours des Carmes

Contrairement au campus de l'université relativement vide et aux bâtiments à l'aspect répétitif, le centre-ville regorge de motifs visuels caractéristiques tels que des enseignes, façades, mobilier urbain, panneaux, et autres éléments pouvant être utilisés comme amers visuels. Par conséquent, il n'a pas été nécessaire d'ajouter, comme dans le premier parcours, certaines cibles artificielles. Celles-ci n'avaient d'ailleurs été ajoutées qu'à des fins d'expérimentation et de validation. En conditions d'utilisation réelles, les deux modules, vision et GPS s'avèrent généralement complémentaires, les zones assez dégagées ayant peu de points visuels caractéristiques bénéficiant en effet d'une bonne couverture GPS, alors qu'à l'inverse les aires urbaines plus denses, souffrant de dégradations des signaux satellites, sont habituellement bien plus riches visuellement, permettant de compenser les imprécisions du GPS par le module vision.

Le long de ce trajet du quartier des Carmes nous avons manuellement créé les modèles Spikenet de 61 cibles à partir d'une vidéo du parcours enregistrée lors d'un premier passage (la liste de ces cibles sera présentée plus tard dans la Figure II-49, elle comprend 20 enseignes, 35 façades, 4 murs, un panneau stop ainsi qu'une boîte aux lettres). Certains de ces amers visuels sont présentés dans la Figure II-46.

Figure II-46 Exemples de points visuels du trajet des Carmes (cibles géolocalisées permettant la correction du GPS).

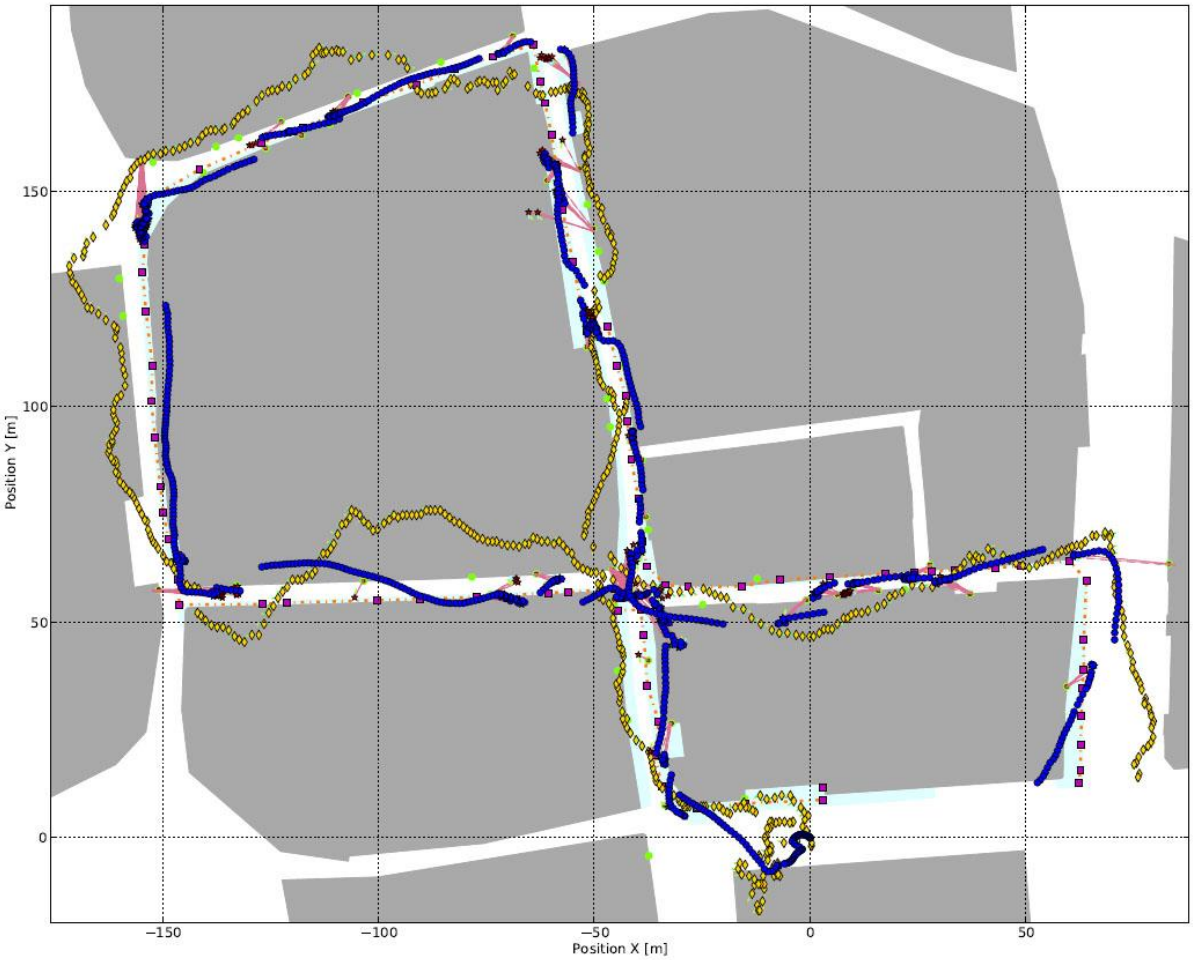


Figure II-47 Relevé du parcours Carmes. Les points en violet correspondent au trajet effectué, ceux en jaune aux positions GPS, et ceux en bleu aux positions résultant de la fusion.

Comme attendu, la précision du GPS sur ce parcours était en moyenne bien plus faible que celle sur le campus en raison des canyons urbains. L'erreur moyenne est en effet centrée autour de 10 m en ville, au lieu de 0 à l'université (bien que l'erreur maximale soit moins importante ici, proche de 20 m contre 30 autour du laboratoire). Si la forme générale de la trajectoire reste assez cohérente avec le trajet effectué, dans une grande majorité des cas la position GPS ne se trouve pas sur le parcours mais dans des bâtiments adjacents (voir Figure II-47). Sur l'ensemble du trajet, l'erreur de positionnement GPS est comprise entre 7 et 12 m dans 56% des cas, et seulement 24% des relevés ont une erreur inférieure à 7 m. Pour les positions fusionnées en revanche 71% sont en dessous de 7 m et 63% en dessous de 4 m, comme illustré dans la Figure II-48.

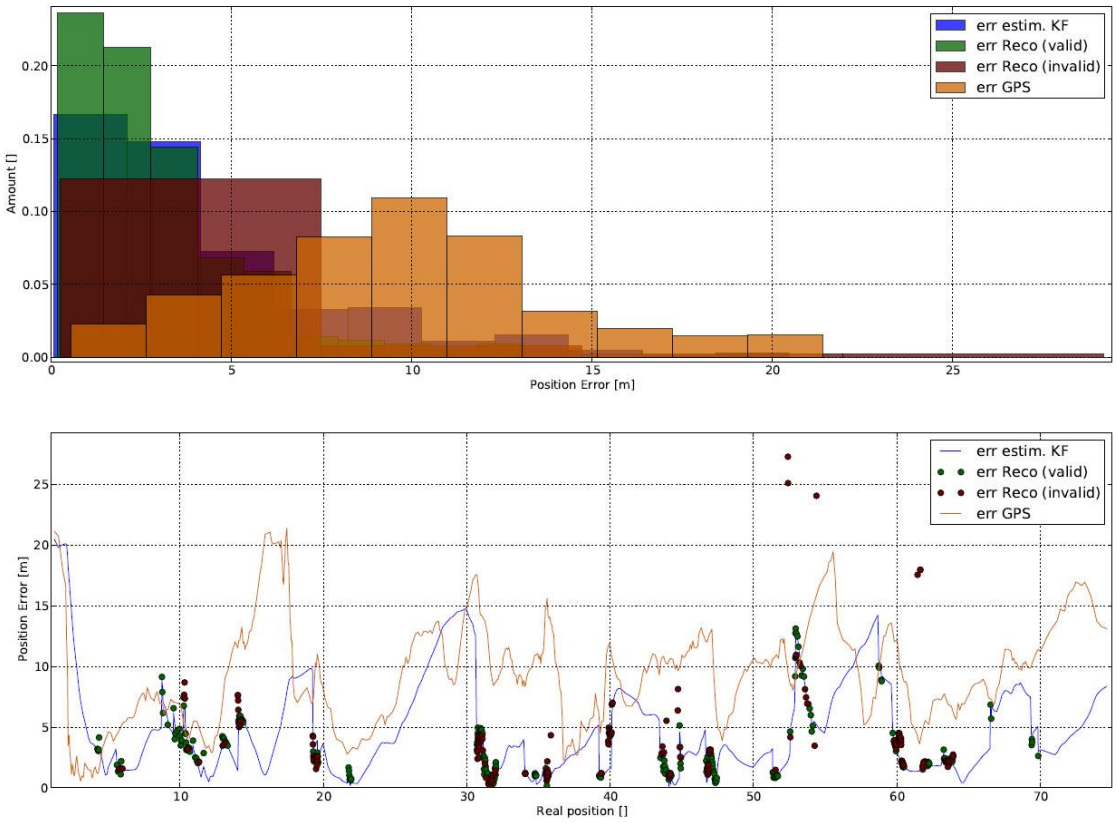


Figure II-48 Distribution des erreurs de positionnement au cours du trajet des Carmes (erreurs du GPS, des détections visuelles correctes, incorrectes, ainsi que de la position fusionnée finale estimée avec un filtre de Kalman).

Du point de vue des reconnaissances visuelles, celles-ci ont bien permis de corriger les imprécisions GPS comme l'ont montré les résultats de l'erreur moyenne de positionnement. Certaines cibles ont pu être reconnues plus aisément que d'autres, comme le montre la Figure II-49. Ceci s'explique en partie par leur position, ou leur taille. En effet, étant donné l'angle de vision large des caméras, des objets petits comme un panneau stop ne pourront être détectés qu'à faible distance, donc pendant une durée bien plus courte qu'une grande façade plus lointaine. De plus les motifs parallèles au sens de la marche (sur le trottoir opposé par exemple) sont très sujets aux déformations perspectives au cours du déplacement, et ne peuvent donc généralement pas être suivis longtemps non plus, à l'inverse des éléments frontaux, perpendiculaires à la trajectoire. Enfin comme les relevés ont été effectués pour certains à plusieurs mois d'intervalle et dans des conditions de luminosité très différentes, les changements visuels induits étaient beaucoup plus importants que sur les cibles artificielles mises en place dans le parcours sur le Campus. Un dernier facteur expliquant les performances légèrement en deçà de celles du premier parcours est l'imprécision du magnétomètre, plus important en centre-ville à cause de sources d'interférence plus nombreuses. Comme nous l'avons expliqué dans la section II.2.4.1, l'erreur d'orientation du casque induit de fortes répercussions dans le calcul de la

position basée sur une détection visuelle. Il arrive donc à quelques reprises durant le trajet qu'une cible soit correctement détectée mais que la position fusionnée résulte pourtant en une localisation incorrecte, non du fait d'une fausse alarme, mais simplement de l'erreur du nord magnétique fourni par la centrale inertielle.

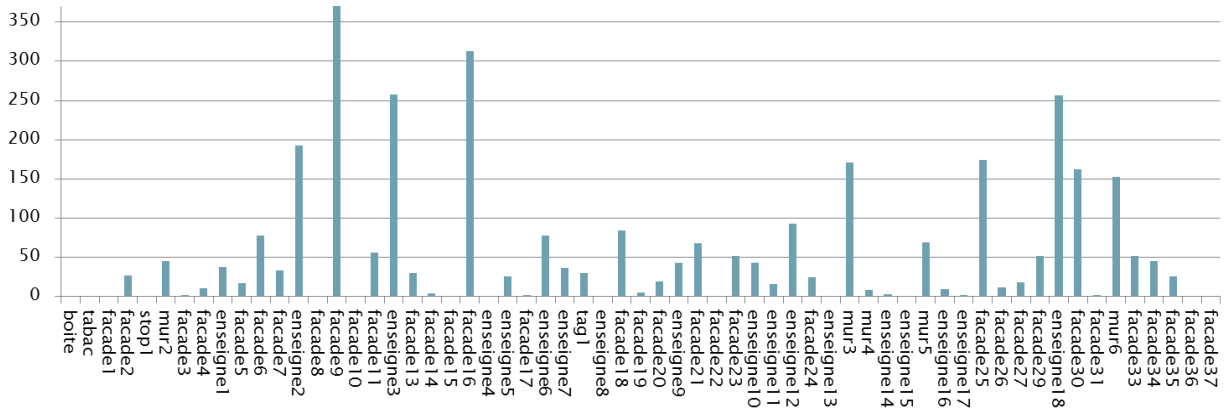


Figure II-49 Nombre de détections de chacune des cibles visuelles du trajet des carmes au cours de 7 répétitions de l'itinéraire.

2.5.3 Expérimentation utilisateur

En plus des expérimentations mentionnées précédemment, destinées à évaluer le fonctionnement des différents sous-modules et plus particulièrement l'influence de la vision sur le positionnement grâce au module de fusion, une expérimentation type « preuve de concept » a été réalisée avec deux utilisateurs déficients visuels de l'Institut des Jeunes Aveugles de Toulouse.

Figure II-50 Trois utilisateurs non-voyants du système Navig.

Le parcours réalisé était celui au sein du campus, aboutissant à l'entrée du laboratoire. Durant la dernière partie de l'itinéraire (une fois arrivé à moins de 20 mètres de la destination finale) le guidage était assuré par la boucle entre vision et son spatialisés décrite plus tôt dans la section II.2.2, grâce aux modèles visuels permettant la détection de la façade et de la porte du bâtiment. Cette fonctionnalité s'est avérée très intuitive et appréciée. Les utilisateurs montraient très peu d'hésitation pour s'orienter et se diriger correctement vers leur but. Différents tests avec des sujets voyants équipés du dispositif ont également montré la facilité d'usage et l'efficacité du système de localisation de cibles en boucle fermée entre la détection et la sonification spatiale.

Le reste du parcours a pu également être effectué (autant par le non-voyant utilisant une canne blanche que celui accompagné d'un chien). Malgré certaines difficultés d'utilisation et quelques problèmes soulevés, ces tests se sont avérés dans l'ensemble assez concluants, et les utilisateurs enthousiastes. Les points à travailler au vu des remarques qui nous ont été faites relevant principalement de l'ergonomie de l'interface (en particulier des métaphores sonores et des instructions de navigation), nous ne développerons pas plus ces questions et proposons de se référer à [Kammoun, 2013]. Nous discuterons en revanche dans la section suivante les aspects perfectibles des composantes visuelles du système et les problèmes rencontrés en lien avec celles-ci.

3. Discussion

Dans ce chapitre nous avons présenté la conception d'un nouveau système de suppléance reposant sur la vision artificielle, permettant l'aide aux non-voyants dans deux tâches particulièrement critiques : la navigation et la localisation d'objets. Nous nous sommes ainsi inscrits dans une approche fonctionnelle et avons notamment pu montrer que le couplage de la reconnaissance d'objets et de la synthèse de sons spatialisés permettait de restaurer des boucles visuomotrices. En outre, la limite de la majorité des aides à l'orientation, à savoir l'imprécision du positionnement par satellites, a également pu être levée grâce à la vision artificielle, par la détection d'amères visuelles permettant de corriger les données GPS.

Néanmoins, différents aspects du système Navig pourraient encore être améliorés, nous proposons donc ci-dessous une discussion des limites constatées et des solutions pouvant être envisagées dans le futur.

3.1 Composantes visuelles

Du point de vue des composantes liées à la vision, qui font l'objet de cette thèse, plusieurs pistes semblent intéressantes à explorer. D'une part dans l'amélioration de l'algorithme de reconnaissance de formes. Il est évident que plus celui-ci sera robuste et rapide, plus les performances globales du système augmenteront. Ceci nous a conduits à développer un nouveau moteur de reconnaissance multi-résolutions, présenté dans le chapitre suivant. Aussi performant soit-il, il restera malgré tout certaines fausses détections occasionnelles et comme nous l'avons vu, celles-ci peuvent influer sur la précision de localisation. La méthode RecoPath, présentée dans la section II.2.4.4, permet de filtrer celles-ci de façon assez efficace. Pour une meilleure identification des fausses alarmes il serait possible d'y ajouter d'autres critères tels que les scores de détection, la taille, la distance et la hauteur de la cible, ou encore le nombre de détections dans une fenêtre temporelle réduite. La fiabilité des détections pourrait de plus être améliorée par un filtrage en sortie basé sur la couleur, qui n'est pas utilisée par Spikenet, mais également par la stabilisation de l'image (en appliquant une rotation correspondant à l'inverse de l'orientation du casque, fournie par la centrale inertielle, de sorte que l'image reste horizontale).

La précision de la stéréovision constitue un autre facteur dont l'amélioration entraînerait un meilleur positionnement lors de la détection d'amers visuels. Depuis une dizaine d'années, de nombreuses méthodes ont été proposées pour améliorer la phase

critique de mise en correspondance. Si jusqu'à peu les techniques d'optimisation globale, coûteuses, fournissaient les meilleurs résultats, de nouvelles approches locales ont récemment vu le jour, dont les performances égalent ou surpassent les précédentes. Les algorithmes globaux reposent sur un problème d'optimisation à l'échelle de l'image complète et certaines hypothèses explicites sur la structure de la scène. Ils visent donc à estimer les valeurs de disparité minimisant une fonction globale de coût, qui combine les données à des termes de *smoothness*[1]. [Wang and Zheng, 2008] proposent par exemple une méthode itérative basée sur la coopération et la compétition entre régions. D'autres algorithmes et leur évaluation sont détaillés dans un état de l'art de Richard Szeliski [Scharstein and Szeliski, 2002]. Les méthodes locales quant à elles centrent une fenêtre sur le pixel de l'image de référence puis déplacent celle-ci le long de la ligne épipolaire de la deuxième vue afin de trouver le point à la correspondance maximale. Déterminer la taille de cette fenêtre est délicat, si trop petite elle risquerait de ne pas capturer suffisamment de variations d'intensité dans les zones relativement uniformes, et si trop grande elle pourrait contenir des points de différentes disparités [Hosni et al., 2009]. De nombreuses techniques ont donc été proposées, avec des fenêtres à taille adaptative, carrées, gaussiennes, voir 3-d pour supporter des surfaces inclinées, se référer à [Scharstein and Szeliski, 2002; Tombari et al., 2008] pour une liste exhaustives des méthodes existantes. Les approches les plus concluantes ayant vu le jour depuis 2005 reposent sur des schémas d'agrégation utilisant une segmentation basée sur la couleur. En assumant que les pixels voisins de même couleur partagent la même disparité, des poids sont calculés pour chacun des pixels de la fenêtre en fonction de leur distance et de leur dissimilarité de couleur par rapport au centre [Gerrits and Bekaert, 2006; Hosni et al., 2009; Tombari et al., 2007; Yang et al., 2009; Yoon and Kweon, 2006], comme illustré dans la Figure II-51.

D'autres techniques, moins courantes, utilisent aussi le couplage à des caméras Time-Of-Flight [Zhu et al., 2008] ou bien l'odométrie visuelle [Skulimowski and Strumillo, 2008], qui consiste à estimer le mouvement des caméras à partir d'une série de points d'intérêt suivis dans plusieurs trames consécutives, afin de corriger la carte de profondeurs en fonction des valeurs précédentes et du déplacement. Quelles que soient les directions prises, ces avancées dans le domaine de la stéréovision se répercuteront de façon bénéfique sur la précision de localisation de la méthode de positionnement que nous proposons. Notons enfin qu'une méthode relativement simple, que nous n'avons pas encore implémentée, permettrait elle aussi de réduire cette erreur. Elle consisterait non pas à prendre le centre de la cible détectée par Spikenet comme référence pour l'estimation de la profondeur si elle était dans une zone uniforme (ce qui impliquerait un risque potentiel

[1] Pourrait se traduire par « continuité », c'est-à-dire que des régions voisines ont de fortes probabilités d'avoir des valeurs de disparité voisines.

d'erreur lors de la mise en correspondance), mais une autre région de la cible dont l'énergie locale serait suffisante.

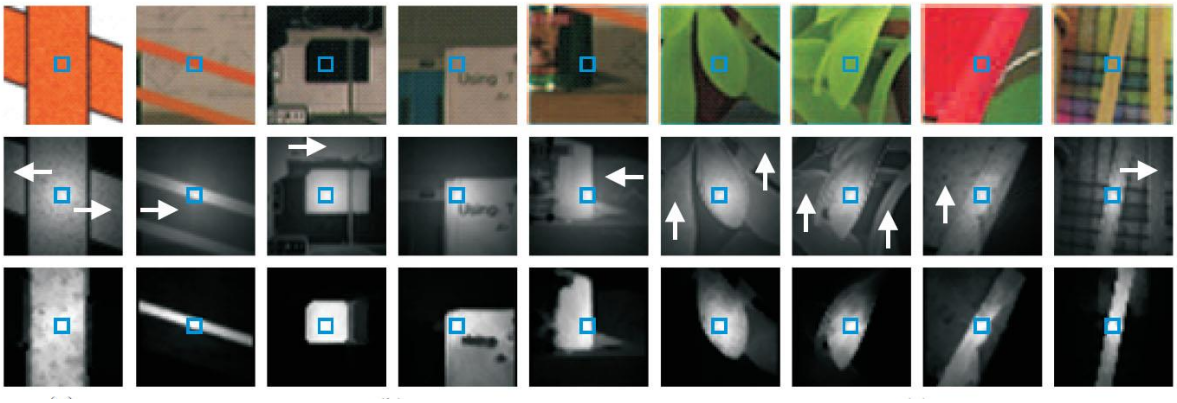


Figure II-51 Régions de support utilisant les informations de couleur sur des patchs tirés de la base Middlebury (1ère ligne). Les résultat de la méthode des poids adaptatifs de [Yoon and Kweon, 2006] sont présentés sur la 2ème ligne, ceux de la distance géodésique de [Hosni et al., 2009] sur la troisième.

Parmi les autres apports et modifications susceptibles d'améliorer les composantes visuelles du système Navig nous pouvons terminer en mentionnant l'utilisation de techniques de suivi (ou *tracking*). Celles-ci pourraient prendre deux formes. Tout d'abord dans l'algorithme de reconnaissance Spikenet, afin d'assurer des détections plus robustes. Lorsqu'un objet a été reconnu avec une confiance suffisante dans une partie de l'image, il semble en effet justifié de penser que celui-ci se trouvera dans l'image suivante à une position voisine et avec des paramètres relativement proches (en taille et en orientation), les trames vidéo étant généralement peu espacées dans le temps (entre 20 et 100 ms selon le *framerate*). Par conséquent les seuils de détection pourraient être adaptés dynamiquement de façon à prendre en compte ces propriétés, en utilisant potentiellement aussi les mouvements de la caméra pour compenser la position prédite de la cible dans l'image en fonction des mouvements de la caméra fournis par la centrale inertielle. Il est également possible d'imaginer un mécanisme de création de modèles à la volée afin d'obtenir un suivi plus robuste aux changements de point de vue et aux déformations perspectives, avec la création et l'activation d'un nouveau modèle dès que la qualité de détection d'une cible suivie passerait en-dessous d'un seuil. Une autre technique de suivi, l'odométrie visuelle, de nature différente, pourrait quand a elle être utilisée pour corriger l'orientation des caméras en cas de mesures incorrectes du magnétomètre, et d'estimer le déplacement relatif de l'utilisateur afin de renforcer l'algorithme de fusion (voir par exemple la méthode proposée dans [Hirota et al., 1996]). En effet, l'odométrie visuelle, notamment utilisée dans les algorithmes de SLAM, consiste à estimer le mouvement de caméras à partir du suivi d'au

moins 3 points fixes de l'environnement. Ceux-ci sont généralement extraits par des détecteurs de points d'intérêt, puis appareillés dans deux images consécutives par corrélation locale, et leurs changements de coordonnées respectives sont finalement utilisés pour estimer les matrices de rotation et de translation des caméras. Ce mouvement pourrait donc être combiné aux méthodes de dead-reckoning basées sur les capteurs inertiels lors de la fusion finale.

3.2 Multi-caméras

Les prototypes du système Navig utilisés jusqu'à maintenant comprenaient, tel que nous l'avons vu, deux caméras frontales. Or, dans la plupart des situations de déplacement d'un piéton, l'orientation de la marche s'aligne sur les trottoirs dans le même axe que la rue, et la majorité des éléments d'intérêt pouvant constituer des amers visuels pour la correction du positionnement (façades, enseignes, magasins, etc…) se trouvent sur les côtés et non face à l'utilisateur. Dans ces conditions, les motifs visuels composant les cibles sont sujets à de fortes déformations perspectives. Avec des caméras latérales en revanche, ces amers resteraient perpendiculaires à l'orientation des caméras, et donc plus aisément reconnaissables.

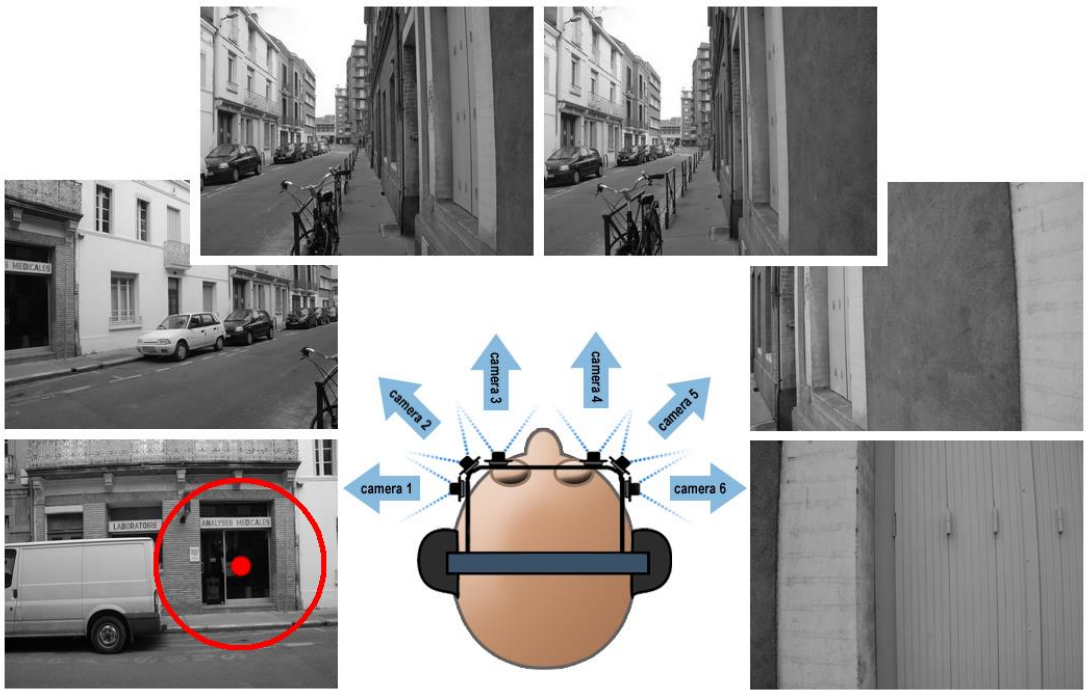


Figure II-52 Concept du prototype 6 caméras.

Ce constat nous a amenés à commencer la mise en place d'un nouveau dispositif utilisant 6 caméras (dont le concept est illustré dans la Figure II-52). Cette architecture, en augmentant au passage le champ visuel, permettra la détection d'un nombre bien plus grand d'amers visuels. Le prototype en question devrait être équipé de microprocesseurs DSP (*Digital Signal Processor*) et de caméras basées sur des capteurs CMOS de 2 millions de pixels, qui couvriront chacune 60° d'angle visuel. Nous travaillons à la mise en place d'un pilote permettant de fournir simultanément deux images de 320x240 px, l'une couvrant l'ensemble du champ visuel, et la deuxième correspondant à un zoom dynamique dans une sous-partie de l'image. C'est donc au total 12 images qui seront reçues à chaque rafraîchissement, 6 images globales à basses résolutions et 6 sous-parties, dans un fonctionnement similaire à la vision humaine, où le champ visuel périphérique permet de détecter des zones d'intérêt et de guider des saccades oculaires venant ensuite effectuer un traitement à plus haute résolution spatiale. Une application de ce procédé est proposée dans la Figure II-53.

Figure II-53 Système à double résolution : les 6 images centrales (colonnes 2 et 3) correspondent aux vues basses résolution de l'ensemble du champ visuel. Dans les zones d'intérêt (déterminées à partir des détections visuelles ou d'autres heuristiques liées à l'énergie locale, à des hypothèses sur la scène ou au suivi de cibles) sont effectués des zooms depuis l'image originale, de sorte à fournir pour chacune une deuxième image de même taille (320x240), qui correspond à une sous-partie de la vue globale. Les zones sélectionnées pour ces zooms sont illustrées par des rectangles blancs dans les images centrales, et les images résultantes sont affichées sur les côtés (colonnes 1 et 4).

Afin de limiter le temps de traitement et réduire les fausses détections, il sera également possible d'utiliser l'orientation de la tête ainsi que le cap de l'utilisateur -calculé par le moteur de fusion- pour rechercher les modèles de façon sélective sur chaque caméra. Par exemple, comme nous connaissons le sens de marche et les coordonnées exactes des cibles à rechercher dans l'environnement, le modèle d'une façade censée se trouver sur la gauche de l'utilisateur ne sera actif que dans le traitement de l'image de la caméra latérale correspondante. Soulignons enfin que certaines modifications du moteur de fusion seront nécessaires pour supporter cette nouvelle configuration. En effet, dans le prototype actuel, la détection d'un seul point visuel suffit à estimer la position de l'utilisateur, celle-ci se trouvant dans le champ de vue stéréoscopique des caméras (on connait donc son orientation et sa distance). Cependant, selon la disposition finale des caméras retenue, il est possible que certains espaces latéraux ne soient couverts que par une caméra. La détection d'un amer visuel dans le champ monoscopique ne fournira donc pas une coordonnée résultante unique (latitude et longitude), mais une droite sur la carte (on contraint les positions sur un plan horizontal, sans chercher à évaluer l'altitude). En revanche, au-delà de deux détections, la situation sera similaire à l'architecture actuelle qui permet d'estimer des coordonnées géodésiques classiques basées sur la vision.

III. Développement d'un algorithme de reconnaissance de formes multi-résolutions

Sommaire de section

1. Introduction

Comme nous l'avons énoncé dans la problématique, l'approche poursuivie au cours de cette thèse consiste à incorporer aux dispositifs d'aides aux non-voyants un traitement de la scène visuelle grâce à des algorithmes de vision artificielle. Par « vision artificielle » nous entendons des méthodes automatiques d'analyse permettant l'extraction d'informations haut-niveau sur l'environnement, telles que la détection et la reconnaissance d'objets. Si cette démarche visant à suppléer la déficience visuelle par des traitements visuels artificiels pourrait sembler naturelle, elle s'avère pourtant très rare dans les systèmes existants à l'heure actuelle. En effet, tel que nous l'avons vu dans l'état de l'art, de nombreux dispositifs ne tirent pas parti de l'information visuelle, mais utilisent plutôt d'autres capteurs tels que des télémètres, radars, accéléromètres, lecteurs RFID ou récepteurs GPS. Parmi ceux incorporant une caméra, plutôt que d'effectuer une réelle analyse de la scène, les traitements de l'image se limitent généralement à un simple sous-échantillonnage, parfois couplé à l'extraction d'arêtes ou à des filtres gaussiens de lissage, qui permettent la réduction du bruit et de la quantité d'informations à transmettre.

Le développement du projet Navig décrit dans le chapitre précédent nous a conduits à nous pencher de plus près sur le moteur de reconnaissance de formes, au centre du système. Il constitue en effet la clé à l'efficacité de l'architecture que nous avons proposée. Les expérimentations de ce dispositif ont révélé plusieurs limites de l'algorithme Spikenet que nous avons utilisé, notamment dans sa tolérance aux déformations perspectives, aux transformations affines, ainsi que dans les temps de traitement nécessaires à la recherche simultanée d'un grand nombre de modèles, ralentissant l'ensemble du système et diminuant ainsi son utilisabilité. Nous nous intéresserons donc dans ce chapitre au domaine de la vision artificielle, en présentant un tour d'horizon des différentes méthodes existantes et en proposant un nouvel algorithme de reconnaissance bio-inspiré adapté à notre contexte d'utilisation.

Les mécanismes biologiques permettant la reconnaissance d'objets, de visages, l'interprétation d'une scène, et l'extraction de nombreuses autres informations haut-niveau relatives à la modalité visuelle, sont des phénomènes d'une très grande efficacité qui sont effectués pour la plupart de façon automatique[1] et dans des délais incroyablement courts [Thorpe et al., 1996]. Pourtant, la nature même de ces traitements reste un grand mystère que tentent de percer les neurosciences visuelles, notamment par l'observation de

[1] C'est-à-dire sans processus attentionnels ou nécessité de raisonnements.

propriétés indirectes ou d'épiphénomènes, et ce afin de proposer des hypothèses sur les aspects computationnels de ces fonctions visuelles.

Indépendamment de cette quête de compréhension du fonctionnement du système visuel humain et animal, s'est développé dans un grand nombre de champs d'application le besoin de traitements informatisés similaires, qui permettent d'extraire algorithmiquement des informations sémantiques variées à partir d'images. La production de contenus multimédia a en effet été multipliée dans des proportions considérables ces vingt dernières années, notamment en ce qui concerne les images et les vidéos (plusieurs milliards de photos indexées par Google ou Flickr, et environ 65 millions de vidéos disponibles sur YouTube). La recherche d'images, et plus généralement de contenus multimédias, est donc un besoin qui va grandissant avec cette disponibilité croissante de documents. L'existence de grandes bases de données d'images annotées permet maintenant d'envisager à large échelle plusieurs directions de recherche combinant la sémantique des images et leur aspect visuel, afin de permettre l'identification plus fine des éléments présents dans la scène.

Nous observons ainsi une convergence des recherches textuelles et basées sur le contenu. Des services de recherche d'images sont maintenant proposés par la plupart des grands moteurs de recherche. Si historiquement ces systèmes n'exploitaient que le texte environnant l'image dans la page web, ils incorporent maintenant quelques caractéristiques propres à l'image, essentiellement sur son format (résolution de l'image, standard vs. panoramique, couleur vs. noir & blanc, photo vs. graphique,...) mais également sur leur contenu (comme la détection de visages désormais intégrée à Exalead et Google Images). L'amélioration de la recherche de documents multimédia fait donc partie des défis actuels, comme en témoignent les nombreux programmes de recherche européens et internationaux, mais aussi les initiatives de groupe privés ("Key Scientific Challenges Program" de Yahoo par exemple, qui finance des projets sur l'extraction d'information et le multimédia). On observe également un nombre croissant de campagnes d'évaluation des systèmes de reconnaissance automatique d'images telles que TrecVID, StarChallenge, Videolympics, PASCAL Visual Object Classes Challenge, imagEVAL, ImageCLEF, VideoCLEF, ImageNet Challenge,... Les résultats issus de ces travaux ont donné lieu à une prolifération de systèmes qui commencent à être mis en œuvre sur des grandes collections telles que Flickr ou ImageNet [Veltkamp and Tanase, 2002]. Si ces méthodes de recherche d'images par le contenu sont majoritairement orientées vers le grand public, celles-ci bénéficient également à certains corps de métiers ayant des attentes plus spécifiques [Eakins et al., 1999], comme par exemple les journalistes et plus généralement le monde de l'édition, ou encore celui de la publicité, de la communication, ainsi que les historiens, les designers,...

En plus de l'indexation d'images par le contenu, de nombreux autres domaines requièrent eux aussi différents types d'interprétation d'images et des algorithmes de vision artificielle adaptés :

- En médecine, pour le diagnostic à partir d'images d'IRM, de radios, ou encore la détection de mélanomes.
- Dans le domaine de la sécurité, pour la reconnaissance de visages ou d'empreintes.
- Dans la surveillance, pour la détection d'intrusions, l'identification de comportements suspects, le suivi de personnes ou de véhicules.
- Dans l'univers des jeux vidéo et du divertissement, où de plus en plus de systèmes utilisent la reconnaissance des gestes.
- Dans le domaine de l'aide au handicap, avec la transcription automatique de la langue des signes, les systèmes de suppléance aux aveugles comme ceux que nous avons présenté le premier chapitre, ou encore le guidage de fauteuils roulants assisté par caméras.

D'autres exemples sont listés dans [Gudivada and Raghavan, 1995], parmi lesquels nous pouvons mentionner la robotique, les systèmes d'information géographique, l'éducation, la défense et les applications militaires, les prévisions météorologiques, etc. En résumé, les champs d'applications sont immenses, et la vision artificielle tend à être partie prenante de notre vie quotidienne avec sa démocratisation dans les smartphones, consoles, et bientôt même dans les automobiles.

Il n'existe évidemment pas de méthode universelle permettant de résoudre ces différents types de problèmes. De très nombreux algorithmes de vision artificielle ont donc été proposés et utilisés selon les besoins du contexte d'application. Si les premiers systèmes de reconnaissance d'objets étaient fondés des approches formelles utilisant des modèles géométriques détaillés [Latecki and Lakämper, 1999; Mehrotra and Gary, 1995; Tirthapura et al., 1998; Van Otterloo, 1988], basés par exemple sur l'alignement de blocs [Roberts, 1963], de cylindres [Binford, 1995, 1971] ou de formes plus complexes [Del Bimbo et al., 1996; Thompson and Mundy, 1987], la tendance est de nos jours aux architectures de traitements hiérarchiques et au recours à l'apprentissage artificiel. On voit aussi l'émergence et l'amélioration de nombreuses approches bio-inspirées, dont les performances arrivent maintenant à rivaliser avec des méthodes plus « ingénieriques ». Un état de l'art de la question est proposé dans ce chapitre, suivi de la présentation de l'élaboration d'un nouveau moteur de reconnaissance de formes bio-inspirée multi-résolutions développé à partir de l'algorithme Spikenet.

2. Vision artificielle

Avec les avancées faites ces dernières années en vision par ordinateur, l'analyse d'images et la reconnaissance d'objets (ou de concepts) commencent à être considérées comme des domaines matures, où de moins en moins de systèmes sont construits *from scratch*. Comme le souligne B. Draper dans [Draper et al., 1999], la plupart des plateformes développées depuis la fin des années 90 sont au contraire mises en place en chaînant différents modules standards de vision utilisant des techniques comme l'*enhancement*, l'extraction d'arêtes ou d'autres primitives, la segmentation de régions, le calcul de flots optiques, l'association de formes, de structures symboliques,... Ces modules sont couplés à d'autres composants non spécifiques au traitement d'images, tels que des classifieurs, des opérateurs de fusion, des techniques d'optimisation de paramètres, etc. Ces différents éléments sont présentés dans cette section.

2.1 Recherche d'image par le contenu

Le terme *Content-Based Image Retrieval (*généralement abrévié CBIR, en français recherche d'image par le contenu) a été introduit pour la première fois en 1992 dans [Kato, 1992] pour décrire un processus de recherche d'images dans une base de données à partir d'éléments liés à la forme et aux couleurs. Il a depuis été très largement utilisé et désigne le fait de sélectionner un ensemble souhaité d'images d'une grande collection sur la base de propriétés (ou *features*) pouvant être extraites automatiquement de celles-ci [Eakins et al., 1999]. Les caractéristiques utilisées pour la recherche peuvent être bas-niveau ou sémantiques, mais le calcul de celles-ci doit être automatisé. Ainsi la recherche dans une base d'images préalablement annotées manuellement ne relève pas du domaine de la recherche d'images par le contenu au sens généralement admis, même si ces annotations se réfèrent effectivement à leur contenu.

La recherche d'images par le contenu tire beaucoup de ses méthodes des domaines du traitement de l'image, et de la vision par ordinateur, ou vision artificielle, dont elle est considérée être un sous-ensemble. Elle se distingue par l'emphase sur la recherche au sein de collections de grande taille à partir de caractéristiques souhaitées, alors que le traitement d'images couvre un spectre beaucoup plus large, incluant par exemple les techniques de compression ou de transmission.

Les premiers systèmes d'indexation et de recherche d'images par le contenu étaient uniquement basés sur une description des documents en terme de caractéristiques bas-niveau (histogrammes de couleur, texture, formes, dimensions, etc.) Ces systèmes permettaient par exemple de rechercher dans une base des images similaires à une image exemple fournie par l'utilisateur, et retournaient donc en sortie non pas une classe d'appartenance, mais un certain nombre d'images jugées pertinentes et similaires à l'image requête proposée. Une autre méthode de recherche dans ces systèmes dits de première génération consistait pour l'utilisateur à formuler sa requête directement à partir de ces caractéristiques bas niveau que nous avons citées (couleur, texture,...).

Le but premier d'un tel système étant de fournir aux utilisateurs des outils efficaces de recherche et de navigation, il est donc nécessaire de prendre en compte les besoins et le comportement d'un utilisateur humain[1]. Or il est difficile pour un individu lambda de formuler une requête en termes de descripteurs bas-niveau. On conviendra qu'il est plus intuitif d'exprimer une attente lors de la recherche d'un document multimédia par un ensemble de mots-clés qu'en terme d'histogrammes de couleur ou de magnitude de gradients. De ces besoins ont émergé les systèmes de deuxième génération.

Cette nouvelle vague de systèmes vise à l'indexation sémantique des images et des vidéos, afin d'offrir à l'utilisateur la possibilité de rechercher des documents à l'aide de concepts ou de mots-clés, comme il en va déjà depuis de nombreuses années en recherche d'information textuelle. Ces concepts sémantiques permettant de décrire le contenu de l'image ou de la vidéo peuvent être de natures diverses, et donc de niveaux d'abstraction différents. Ils peuvent aussi bien représenter un objet que des lieux, des actions, des sujets thématiques, ou bien encore des personnes. Une classification de ces requêtes en trois niveau a été introduite dans [Eakins, 1998, 1996].

- Le niveau 1 comprend les recherches basées sur des descripteurs bas niveau tels que les couleurs, textures, formes, positions spatiales, décrites comme *primitive features* dans [Gudivada and Raghavan, 1995].
- Le deuxième niveau correspond aux requêtes de contenu sémantique. Il nécessite l'extraction d'attributs haut-niveaux (nommés logicial features par Gudivada et Raghavan) tels que la présence ou l'identité d'objets dans l'image.
- Le troisième niveau enfin, relève de notions beaucoup plus abstraites, parfois même subjectives, impliquant des traitements complexes de la scène, des objets, des personnes, pouvant nécessiter des processus de raisonnement, d'interprétation, de

[1] On parle souvent de systèmes Human-Centered [Jaimes and Sebe, 2007]

déduction. Ceux-ci interviennent par exemple pour l'identification d'émotions, d'actions et d'événements.

La grande difficulté dans le cadre de la recherche d'images est l'extraction de ces informations "sémantiques". Dans le domaine textuel il est possible de rechercher directement un concept au sein d'un document, et ainsi de pouvoir renvoyer à l'utilisateur formulant par exemple la requête "voiture" les documents contenant ce mot. En revanche comment déterminer si une image contient bien une voiture si l'on ne dispose pas de métadonnées sur celle-ci (annotations ou texte entourant l'image) ? C'est là tout l'enjeu de la recherche par le contenu dans des bases d'images ou de vidéos. Cette distance entre d'une part le signal brut de l'image (un simple vecteur d'intensités de pixels), et de l'autre les concepts sémantiques qui la définissent, est appelée le *fossé sémantique*. Ce terme introduit par Eakins est défini ainsi dans [Eakins et al., 1999] :

> *Le franchissement du fossé sémantique consiste à inférer des caractéristiques haut-niveau, nécessitant un certain degré de raisonnement logique, à partir des informations primitives qu'une machine est capable d'extraire d'une image, telle que ses couleurs ou sa texture.*

2.2 Classification d'images

La classification d'images est une tâche qui permet de déterminer par exemple la présence d'objets (avions, chaises, voitures, etc.), d'événements (manifestations, tremblements de terre), ou encore de scènes (environnement urbain, studio, plage,...). La grande majorité des systèmes de recherche d'images [Gemert et al., 2006; Jianguo Zhang et al., 2006] ou de vidéos [Chang et al., 2007; Smeaton et al., 2008] utilise pour cela une méthode d'apprentissage supervisé consistant à entraîner des classifieurs à partir de données extraites des documents multimédia par des descripteurs. Lorsque différentes sources d'informations sont utilisées, dans le cas par exemple où plusieurs descripteurs sont appliqués aux images ou vidéos, une dernière étape consistant à combiner ces différentes informations par des opérateurs de fusion est alors nécessaire. Le pipeline communément utilisé est résumé dans la Figure III-1.

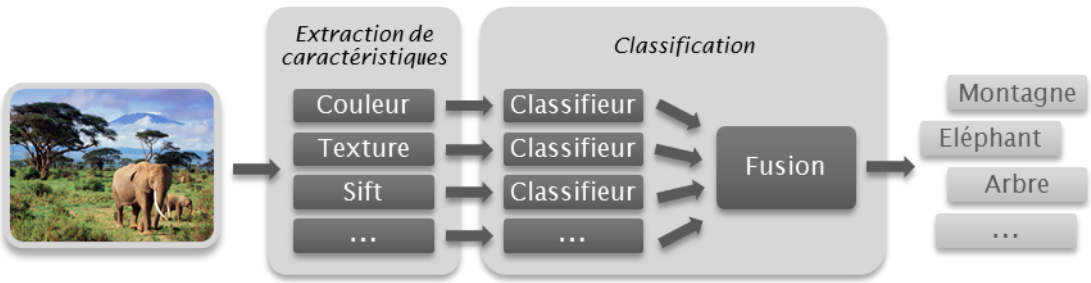


Figure III-1 Architecture standard des systèmes d'indexation d'images par le contenu.

La plupart des systèmes de détection de concepts dans des bases d'images ou de vidéos sont des combinaisons des descripteurs et classifieurs présentés dans les sections suivantes. On observe depuis quelques années une forte convergence des systèmes proposés vers une architecture *mainstream* consistant en un pipeline descripteur-classifieur-fusion. Au niveau des classifieurs utilisés on observe de moins en moins de diversité, les équipes bénéficiant de l'expérience des recherches conduites dans la communauté CBIR et des nombreuses études comparatives qui ont été menées. La plupart des systèmes optent pour des SVM, et s'appuient dans leur implémentation sur des librairies matures telle que la très populaire libSVM [Lin and Chang, 2001], ou la librairie SVMlight [Joachims, 1999], également répandue. Les paramètres les plus courants de ces SVM sont des noyaux RBF optimisés par une méthode grid-search 3 ou 5-fold, et des tirages aléatoires pour la sélection des exemples négatifs. Cette fonction kernel (*Radial Basis Function*), a montré de meilleures performances dans la plupart des études, cependant pour les approches à base de *codebook* (Bag of Visual Words), Zhang et al. ont suggéré que *l'Earth Mover Distance* ou le noyau χ^2 devaient être préférés [Jianguo Zhang et al., 2006]. Au niveau des descripteurs, si les histogrammes de couleur, filtres de Gabor ou moments de couleur restent présents comme source additionnelle d'information, l'attention se focalise maintenant sur les SIFT et ses dérivés, couplés à des représentations de type *Bag-Of-Visual-Word* (ou BOVW [Yang et al., 2007]. Les BOVW consistent à regrouper par des méthodes de clustering les différents vecteurs extraits par le descripteur local pour chacun des points d'intérêt. On considère alors ces différents clusters de points d'intérêt comme des mots visuels, et d'une façon similaire à la recherche textuelle on tentera de prédire la classe d'appartenance d'une image à partir des relations entre les mots visuels et les différents concepts appris. Le nombre de clusters choisi définit la taille du *codebook*, c'est à dire le nombre de mots visuels. A partir de cette seule approche, avec un descripteur correctement paramétré (taille du *codebook*, choix et ajustement des détecteurs de points d'intérêt) les équipes Surrey, Mediamill et Columbia ont montré des résultats surpassant la plupart des autres Systèmes.

2.3 Descripteurs

L'extraction de caractéristiques constitue le premier pas de toutes les procédures d'analyse d'images qui visent à un traitement symbolique de leur contenu. Les descripteurs bas-niveau fournissent en effet une première représentation symbolique à partir du signal brut de l'image (les pixels), et constituent donc une étape majeure dans le franchissement du fossé sémantique. Les éléments de base de la plupart des descriptions symboliques d'images sont les points, les arêtes et les régions [Förstner, 1994]. De très nombreuses méthodes existent pour décrire l'image en termes de caractéristiques bas niveau, constituant autant d'angles d'interprétations possibles du contenu de l'image. Parmi les

systèmes de reconnaissance de concepts, le nombre de ces descripteurs utilisés peut aller de un jusqu'à plus d'un millier. Ceux-ci se réfèrent généralement à la couleur, à la texture, à la forme ou à d'autres propriétés spatiales. Ils peuvent être appliqués sur toute l'image ou seulement sur certaines parties, et traités indépendamment ou non de l'échelle ou de l'orientation [Aigrain et al., 1996]. Les principales approches sont résumées dans cette partie.

2.3.1 Descripteurs globaux

Une description statistique globale des caractéristiques d'une image est une technique très utilisée dans l'analyse d'images pour l'indexation et la recherche de documents. Ces attributs globaux sont facilement calculables et réussissent souvent à capturer une information pertinente sur le contenu de l'image. Bien qu'ils ne permettent pas d'évaluer la distribution spatiale des caractéristiques de l'image et donc de sa structure interne, leur importance ne doit pas être sous-estimée. Selon le contexte leur aide peut être précieuse, rechercher les images contenant un grand nombre de lignes droites est par exemple un bon critère dans la détection des constructions humaines. De manière générale, les descripteurs globaux peuvent apporter d'importants indices sur l'apparence visuelle globale d'une image, son type, ou certaines autres propriétés. Nous avons regroupé les principaux descripteurs globaux en deux grandes classes détaillées ci-dessous, ceux permettant de représenter les informations de couleur d'une part, et ceux décrivant la texture de l'image.

2.3.2 Couleur

L'analyse des couleurs est une des premières méthodes utilisées pour la recherche d'images et les tests de similarité. Celle-ci est non seulement aisée et rapide, mais offre une robustesse aux changement de lumière, d'angle de vue et d'échelle supérieure à beaucoup d'autres techniques [Aigrain et al., 1996]. Il existe de nombreuses façons de modéliser cette information mais la plus répandue est l'utilisation d'histogrammes de couleurs, introduits par [Swain and Ballard, 1990], qui fournissent la distribution globale des couleurs dans l'image. Plusieurs variantes existent, dépendant notamment du choix de l'espace de couleurs utilisées.

On peut ainsi citer les histogrammes RGB, qui sont une combinaison de trois histogrammes à une dimension calculés pour chacune des composantes rouge, verte et

bleue. En utilisant l'espace colorimétrique HSV[1], qui donne la teinte, la saturation et la valeur (brillance) de chaque couleur, on obtient des histogrammes de teinte (ou Hue Histogram), également à 3 dimensions, mais invariants aux changements d'intensité de la lumière. Plusieurs autres espaces de couleurs peuvent être utilisés, comme YCrCb[2], Color-Oppoent, RG[3], ou HMMD[4]. Se référer aux articles [van de Sande et al., 2008] et [Manjunath et al., 2001] pour un descriptif plus complet de ces méthodes et l'étude de leurs propriétés d'invariance aux changements d'intensité et de couleur de la lumière.

Les histogrammes de couleurs ne modélisant pas la distribution spatiale des couleurs, une image contenant une grande région rouge sur un fond vert aura le même histogramme qu'une autre ayant le même nombre de pixels rouges et verts aléatoirement répartis au sein de l'image. Les moments de couleurs [Stricker and Orengo, 1995; Stricker and Dimai, 1996] sont une alternative permettant d'incorporer, à différents degrés, des informations sur la répartition spatiale des couleurs. L'idée derrière cette approche est que toute distribution de couleur peut être caractérisée par ses moments. De plus, comme les informations les plus importantes sont concentrées dans les premiers moments, le descripteur peut se contenter d'extraire les moments de premier ordre (moyenne), deuxième et troisième ordre (variance et asymétrie). Les moments de couleur généralisés M_{pq}^{abc} d'ordre $p+q$ et de degré $a+b+c$ ont été définis dans [Mindru et al., 2004] par la formule suivante, où I est la fonction associant à chaque point de l'image de coordonnées (x,y) la valeur du pixel pour chacune des composantes de couleur ($I : (x,y) \rightarrow (R(x,y), G(x,y), B(x,y))$) :

$$M_{pq}^{abc} = \iint x^p y^q \left[I_{R(x,y)}\right]^a \left[I_{G(x,y)}\right]^b \left[I_{B(x,y)}\right]^c dxdy$$

En plus des histogrammes de couleurs et des moments de couleurs on peut également mentionner quelques autres descripteurs souvent abordés dans la littérature : les *Color Sets* [Smith and Chang, 1996a], les *Color Coherence Vector* [Pass and Zabih, 1996], les histogrammes de corrélation de couleur[5] [Huang et al., 2001] ou encore les descripteurs SCD (*Scalable Color Descriptor*) [Manjunath et al., 2001], *Dominent Color* [Deng et al., 2001] et CSD (*Color Structure Descriptor*) [Manjunath et al., 2001].

[1] Hue Saturation Value.
[2] Que l'on désigne également sous le nom YUV, décrivant les couleurs en terme de luminance et de chrominance.
[3] Normalized RGB : *en normalisant la somme des 3 composantes RGB, seules les deux premières sont alors nécessaires, la troisième pouvant être calculée par* $B = 1 - R - G$ *(cette normalisation procure l'invariance aux changements d'intensité).*
[4] Hue-Min-Max-Difference.
[5] Color Correlograms.

2.3.3 Texture

La notion de texture, bien qu'en général comprise de tous, reste très difficile à définir de manière formelle, aucune des différentes définitions proposées n'a réussi à faire consensus [Tuceryan and Jain, 1993]. Certains ont préféré la définir par ce qu'elle n'est pas : *"la texture est tout ce qui reste après avoir considéré la couleur et les formes présentes dans l'image"*. Plus spécifiquement, elle décrit la structure au niveau macroscopique, elle caractérise les motifs visuels définis par leur arrangement dans l'image. Une manière plus simple d'appréhender cette notion est d'en donner quelques exemples.

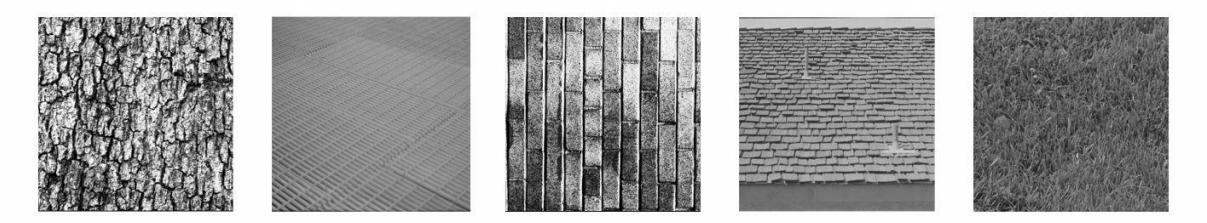


Figure III-2 Exemples de différentes textures

De nombreuses propriétés permettent de caractériser une texture, grandement inspirées par des études perceptives chez l'humain. Parmi elles on trouve le contraste, la granularité, la directivité, la périodicité, la régularité, ou encore l'entropie (se reporter à [Tamura et al., 1978] et [Liu and Picard, 1996] pour une définition de ces différents termes et leur formalisation en terme de traitement d'image). Dans les années 70, Hararalick et al. ont proposé une des premières méthodes de caractérisation baptisée *Co-Occurrence Matrix of Texture Feature* [Haralick et al., 1973]. Cette approche explorait les dépendances spatiales des textures en construisant d'abord une matrice de co-occurrence basée sur l'orientation et la distance entre les pixels de l'image puis en représentant la texture par l'extraction de statistiques sur cette matrice. Des études ultérieures ont montré que les mesures statistiques au plus fort pouvoir discriminant étaient le contraste, l'entropie et la différence inverse des moments [Gotlieb and Kreyszig, 1990].

De nos jours les filtres de Gabor [Bovik et al., 1990] sont souvent reconnus comme les descripteurs les plus efficaces pour représenter textures et surfaces [Rui et al., 1997; Smith and Chang, 1996b]. Ils permettent la détection de contours et motifs selon différentes orientations et échelles. L'image est découpée en blocs réguliers, pour lesquels sont calculées les moyennes et déviations standard de l'énergie des pixels, généralement pour 5 échelles et 8 orientations différentes. D'autres approches assez similaires que nous ne détaillerons pas sont semblables aux transformées de Gabor, notamment les décompositions à base d'ondelettes, les transformées de Fourrier ou encore les transformées de Hough.

Les *Edge Orientation Histograms* [Park et al., 2000] sont d'autres descripteurs permettant de capturer la distribution spatiale des contours. Cette distribution constitue une bonne signature de texture. Leur calcul consiste à diviser l'image en 4×4 blocs, dans lesquels sont calculés des histogrammes locaux pour 5 types d'orientation (0°, 45°, 90°, 135° et isotropiques, présentés ci-contre).

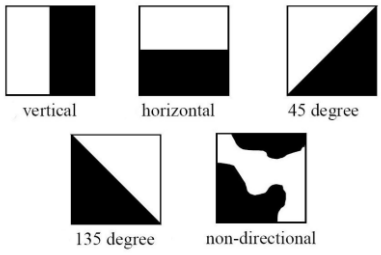


Figure III-3 *Edge Orientation Histograms*

Une autre représentation de la texture de l'image, basée sur une analyse multi-résolution des niveaux de gris, est proposée par le descripteur *Local Binary Patterns* (ou LBP) [Ojala et al., 2000]. Ses valeurs sont calculées sur des régions de 3x3 pixels autour de chaque point de l'image, en comparant les différences d'intensité entre le centre et les 8 pixels de son voisinage. Ainsi pour le point p de coordonnées (xp, yp) la valeur du descripteur est donnée par $LBP(x_p, y_p) = \sum_{n=1}^{8} 2^n \times signe(In - Ip)$ où Ip correspond à l'intensité du point p et In à celles de son voisinage parcouru comme indiqué dans la figure ci-contre. Ce descripteur est très populaire pour sa tolérance aux changements d'illumination, sa rapidité de calcul et son invariance aux rotations (l'invariance aux rotations a été obtenu par une extension des LBP, le *Generalized Local Binary Pattern Operator* [Ojala et al., 2001]).

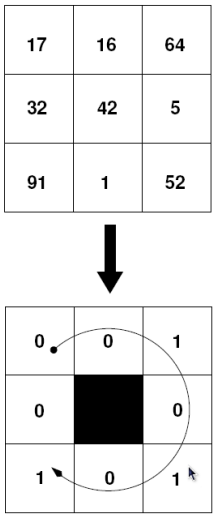


Figure III-4 *LBP*

Parmi les autres descripteurs de texture, signalons enfin les modèles markoviens (*Markov Random Field Representation* [Cross and Jain, 1981, p. 81]), le filtrage multi-source, ou encore les représentations basées sur les fractales (*Fractal-Based Descriptors* [Pentland, 1984], *Fractal Features* [Kaplan et al., 1997]).

2.3.4 Descripteurs de région

On parle souvent de descripteurs de régions lorsque l'on applique des descripteurs "globaux" sur un sous-échantillonnage de l'image. Ils traitent donc au final l'image originale de façon "locale", mais sont à différencier des descripteurs locaux que nous aborderons dans la section suivante, qui font généralement référence aux approches basées sur la détection de points d'intérêt dans l'image. Pour les descripteurs de régions[1], l'analyse "locale" est le

[1] On parle également de descripteurs par "patchs".

résultat d'une division de l'image en blocs réguliers, généralement caractérisée par 2 paramètres : le nombre de patchs en X et en Y, qui peuvent aussi dans certains cas être complétés par la taille des patchs pour spécifier le pourcentage de recouvrement.

Les descripteurs globaux sont alors appliqués sur chacune des régions de manière indépendante (pour le descripteur, chaque bloc est traité comme une image différente), puis les vecteurs de caractéristiques extraits de chaque patch sont ensuite concaténés en un vecteur final. A titre d'exemple si on prend le cas d'un histogramme de couleur RGB classique de dimension 3, après avoir divisé l'image en 8×6 régions et calculé l'histogramme de chacun des 48 blocs, on obtiendrait au final un vecteur de caractéristiques à 144 dimensions ($nbPatchesX \times nbPatchesY \times dim(GlobalDescriptor)$).

2.3.5 Descripteurs locaux

Comme nous l'avons expliqué les descripteurs d'images globaux, bien que performants pour une grande quantité de requêtes, ont néanmoins certaines limitations. Ils sont beaucoup moins efficaces lorsque, par exemple, on recherche des images d'un concept donné prises dans des conditions de vue très différentes. La distribution globale des couleurs et textures change alors radicalement d'une image à une autre, et une description plus locale est alors requise, pour capturer la structure interne de l'image, d'une manière robuste aux changements de point de vue ou d'illumination.

Les descripteurs locaux apportent souvent cette robustesse aux occlusions et aux changements de conditions de vue (lumière, bruit, point de vue), ainsi que l'invariance aux rotations ou aux changements d'échelle. Ils sont communément considérés comme les méthodes offrant les meilleures performances et de plus ne nécessitent pas de segmentation de l'image. Ces descripteurs locaux sont généralement calculés en deux étapes. La première consiste à identifier des points d'intérêt dans l'image, puis les valeurs du descripteur sont ensuite produites en utilisant les caractéristiques de l'image autour de chacun de ces points d'intérêt précédemment trouvés. La détection de ces points est généralement effectuée à différentes échelles, obtenues en sous-échantillonnant l'image initiale.

Il existe une grande variété de descripteurs locaux, qui diffèrent par exemple dans la façon de calculer les points d'intérêt, tâche des détecteurs, tels que les détecteurs de Laplace, Hariss, Susan, *Laplacian of Gaussian*, Forstner,.... L'article [Mikolajczyk and Schmid, 2004] propose une analyse détaillée des plus couramment utilisés et de leurs propriétés.

En plus des différences dans le choix du détecteur les descripteurs locaux se distinguent également par le nombre de points retenus et la façon de traiter l'image autour

de ceux-ci. Le rôle du descripteur est de caractériser l'apparence locale de l'image autour des points identifiés. De ces différents paramètres dépendent les propriétés d'invariance listées précédemment : l'invariance au changement d'échelle, à la rotation, à la translation ou au changement de point de vue [Lejsek et al., 2006]. De plus on peut obtenir la robustesse au bruit par un filtrage passe-bas, et enfin la tolérance aux variations d'éclairage en considérant les dérivées de l'image plutôt que les valeurs brutes de couleurs ou de niveaux de gris. Nous présenterons ici certains des descripteurs locaux les plus courants, une liste plus exhaustive et de plus amples détails sont disponibles dans les articles suivant : [Li and Allinson, 2008; Mikolajczyk and Schmid, 2005, 2004; Moreels and Perona, 2007; Roth and Winter, 2008].

- **SIFT** [Lowe, 2004] : Ce descripteur utilise des histogrammes contenant la position et l'orientation de gradients autour des points d'intérêt. La position est modélisée par une grille 4×4, pour laquelle chaque secteur contient l'angle des gradients selon 8 orientations. La dimension du vecteur de caractéristiques est donc de 128. L'invariance du descripteur est obtenue en assignant à chaque point d'intérêt différentes orientations basées sur les directions des gradients au niveau local. D'autre part la normalisation finale du vecteur par la somme de ses composantes assure l'invariance aux changements d'illumination.

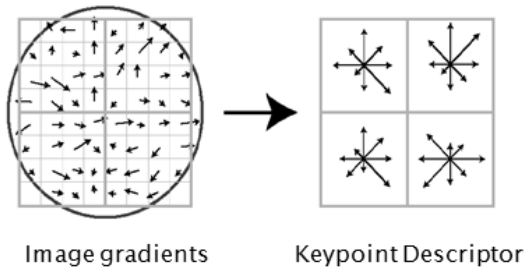


Figure III-5 Description of SIFT (tirée de (Lowe, 2004)) : la magnitude et l'orientation de gradients est d'abord calculés en chaque point dans la région entourant le point d'intérêt. Ceux-ci sont pondérés par une fenêtre gaussienne représentée par le cercle, puis ajoutés à des histogrammes couvrant 4x4 sous-régions, la norme de chaque vecteur correspondant à la somme des magnitudes de gradients de même orientation dans la région. considérée.

- **HOG** [Dalal and Triggs, 2005] : Les histogrammes de gradients orientés sont un dérivé des SIFT se différenciant principalement par une grille beaucoup plus dense avec un fort recouvrement, et des méthodes sophistiquées de normalisation du contraste local au sein des blocs qui se chevauchent .
- **PCA-SIFT** [Yan Ke and Sukthankar, 2004] : L'algorithme PCA-SIFT modélise les gradients de l'image au voisinage des points d'intérêt de façon similaire aux SIFT. Les patchs locaux sont cependant réduits à un espace de dimension inférieure en utilisant une analyse en composantes principales (*Principal Component Analysis*). Ce descripteur est un vecteur de gradients calculés selon les directions X et Y au niveau

des régions de support. Ces régions sont échantillonnées en grilles de 39×39 secteurs, qui produisent donc un vecteur de dimension 3042, réduite à 36 par l'analyse en composantes principales. Comparé aux SIFT ce descripteur a montré de meilleures performances en reconnaissance, et offre de plus un gain en termes de rapidité de par sa représentation compacte.

- **GLOH** [Mikolajczyk and Schmid, 2005] : *Gradient Location-Orientation Histogram* est également une extension des descripteurs SIFT, permettant d'accroître leur robustesse. Le descripteur est calculé pour une grille log-polaire, divisée en 8 angles, et utilisant chacun 3 rayons différents (6, 11 et 15), produisant ainsi 27 bins. Les orientations de gradients sont quantifiées en 16 valeurs, conduisant à un histogramme à 272 bins, dont la dimension est enfin réduite comme dans le descripteur précédent par une analyse en composantes principales.
- **Steerable Filters** [Freeman and Adelson, 1991] et **Differential Invariants** [Schmid and Mohr, 1997] : Les *Steerable Filters* sont des filtres orientés, calculés à partir d'une combinaison linéaire d'un ensemble de filtres de base construits en utilisant des dérivées gaussiennes calculées selon différentes directions, assurant ainsi l'invariance aux rotations. Le descripteur *Differential Invariants* utilise une approche similaire en calculant les dérivées locales de l'intensité de l'image jusqu'au 3ème degré.
- **Shape Context Descriptor** [Belongie et al., 2002, 2006] : Cette dernière technique est basée sur les contours. Ils sont extraits par un filtre de Canny puis leurs positions et orientations, en coordonnées log-polaires, sont quantifiées en histogrammes. On obtient au final un vecteur de dimension 36 qui constitue une bonne façon de décrire les formes et de mesurer leur similarité.

2.3.6 Forme

La forme est comme la texture une notion vague regroupant de nombreux aspects et dont il n'existe pas de définition universelle. Elle peut s'exprimer par la couleur, des motifs, des textures, ou différents autres éléments à partir desquels il est possible de dériver une représentation géométrique [Veltkamp and Tanase, 2002]. Les descripteurs de forme sont donc variés. Certains reposent sur des caractéristiques globales comme l'aspect-ratio, la circularité, ou les moments algébriques [Niblack et al., 1993; Prokop and Reeves, 1992]. D'autres s'intéressent aux contours des formes plutôt qu'à leur surface, par le biais de descripteurs locaux traduisant des segments consécutifs articulés [Mehrotra and Gary, 1995], ou de coefficients de Fourrier comme proposé dans [Van Otterloo, 1988]. Un autre moyen couramment employé pour représenter les contours est l'utilisation de *turning angle*

functions [Latecki and Lakämper, 1999]. Parmi les autres méthodes alternatives de comparaison de formes nous pouvons mentionner les approches à base de déformations d'objets prototypiques [Del Bimbo et al., 1996; Sclaroff and Pentland, 1995], ou encore les représentations sous forme de squelettes utilisant les techniques d'appariement de graphes [Sebastian et al., 2001; Tirthapura et al., 1998].

Néanmoins ces différents descripteurs ne semblent pas constituer de bons prédicteurs d'un jugement humain sur les similarités de formes [Mumford, 1991; Scassellati et al., 1994]. Et s'ils ont été dans les années 60 une des premières pistes dans la reconnaissance d'objets en se basant sur la géométrie computationnelle, ces approches formelles sont de moins en moins utilisées [Mundy, 2006]. Nous ne détaillerons donc pas plus ces méthodes et proposerons pour un état de l'art exhaustif de se reporter à [Mehtre et al., 1997; Veltkamp and Hagedoorn, 2001; Zhang and Lu, 2004]. Notons par ailleurs que nous nous sommes ici focalisés sur les aspects géométriques, mais que beaucoup des descripteurs locaux présentés précédemment, ainsi que des descripteurs de textures ou de couleurs encodent eux aussi des propriétés liés à la notion de formes.

2.4 Classifieurs

Afin de déterminer les caractéristiques visuelles des classes d'objets que nous cherchons à détecter, il est nécessaire d'avoir un ensemble d'images annotées, c'est à dire une base de référence pour laquelle nous est donnée la présence ou l'absence des concepts dans chacun des documents, qui nous permettra d'apprendre ces classes à partir de leurs exemples. Ceci nous positionne donc dans une démarche d'apprentissage supervisé, par opposition à l'apprentissage non-supervisé qui consiste à regrouper un ensemble de données non annotées en groupes homogènes, ceux-ci ne donnant par ailleurs aucune information sur leur interprétation "sémantique".

En apprentissage supervisé, la classification consiste à estimer une fonction $y = f(X)$ à partir d'un ensemble d'exemples de la forme $\{(X_1, y_1), \cdots, (X_n, y_n)\}$, les valeurs de y appartenant à ensemble fini de classes $\{1, \cdots, K\}$, représentant dans notre cas les concepts à détecter. La fonction apprise est appelée un classifieur. L'apprentissage est donc réalisé sur un ensemble de couples (X_i, y_i) exprimant la présence du concept $y_i \in \{1, \cdots, K\}$ dans le i^{eme} document de la base d'apprentissage. Les variables X_i sont typiquement des vecteurs de la forme $\langle x_{i,1}, x_{i,2}, \cdots, x_{i,dim} \rangle$ dont les composantes sont des valeurs réelles caractérisant l'échantillon i (appelées *features* ou *caractéristiques* de X_i). Ces vecteurs sont généralement extraits par les descripteurs bas-niveau, dont nous avons détaillé le rôle et le fonctionnement dans la section précédente.

Les classifieurs, comme nous venons de le voir, produisent donc pour chaque nouvelle instance un label, correspondant à la classe prédite, mais ils peuvent également fournir une estimation de la fiabilité de cette prédiction. On parle alors de *confiance*. P. Fabiani propose une analyse de la pertinence[1] et de la confiance dans [Fabiani, 1996] :

> *La pertinence d'une information ou d'un état de croyance est toujours définie par référence à l'état effectivement réalisé dans le monde réel. Cette référence au monde réel souligne la difficulté de gérer cette pertinence : on peut tout au plus en faire une estimation. Nous préférons donc parler de confiance d'un système par rapport à une information ou un état de croyance incertain. Cela nous donne la définition suivante, par référence à la pertinence.*
>
> *La confiance d'un système envers une information ou état de croyance incertain est l'estimation qu'il peut faire de la pertinence de cette information ou état de croyance incertain par rapport à l'état effectivement réalisé de l'environnement ou monde réel.*

Cette définition générale de la confiance en une information peut s'appliquer au problème de classification d'images. En effet la plupart des classifieurs binaires produisent en plus d'un jugement (présence ou absence du concept dans l'image) un score ou une probabilité qui peuvent être interprétés comme une mesure de confiance de la prédiction par rapport à l'état réel qu'il tente d'estimer par les modèles qu'il a appris.

Les classifieurs SVM produisent directement ce score pour chaque instance évaluée, qu'il est alors aisé de ramener à une décision $\{-;+\}$ en seuillant le score retourné (si $confiance > 0.5$ alors $+$, sinon $-$). Pour d'autres types de classifieurs ne fournissant pas directement cette mesure il est généralement possible de l'inférer de différentes façons. Pour un réseau bayésien naïf binaire par exemple, on peut estimer la confiance en calculant le ratio :

$$f(x) = log \frac{\mathrm{p}(+|\mathrm{x})}{\mathrm{p}(-|\mathrm{x})}$$

La présence du concept sera considérée comme vraie seulement dans le cas où $p(+|x) > p(-|x)$, c'est à dire si $f(x) > 0$. Le signe de $f(x)$ donne donc le label prédit, et la valeur absolue $|f(x)|$ peut être interprétée comme la confiance. Cette mesure de confiance est essentielle dans le contexte de la recherche d'images, car il ne s'agit pas uniquement de déterminer si un concept est présent ou non dans un document, mais de retourner à l'utilisateur une liste ordonnée des exemples du concept cherché, des plus probables aux moins probables.

[1] Fiabilité d'une information incertaine.

2.4.1 Support Vector Machines

Les SVM[1] sont des méthodes de classification binaire par apprentissage supervisé qui furent introduites par Vapnik en 1995 [Vapnik, 1995]. Ces méthodes conçues pour une séparation de deux ensembles de données reposent sur l'existence d'un classificateur linéaire dans un espace approprié.

Pour des données linéairement séparables il existe une infinité d'hyperplans séparant sans erreur ces ensembles. L'hyperplan optimal, appelé *hyperplan à marge maximale* est celui situé à la distance maximale des vecteurs les plus proches parmi la base d'exemples (voir Figure III-6). Le but des SVMs est de trouver cet hyperplan maximisant la marge de séparation entre deux classes. Nous ne détaillerons pas ici les algorithmes d'optimisation de la recherche des hyperplans, qui sont fournis dans le livre de Vapnik.

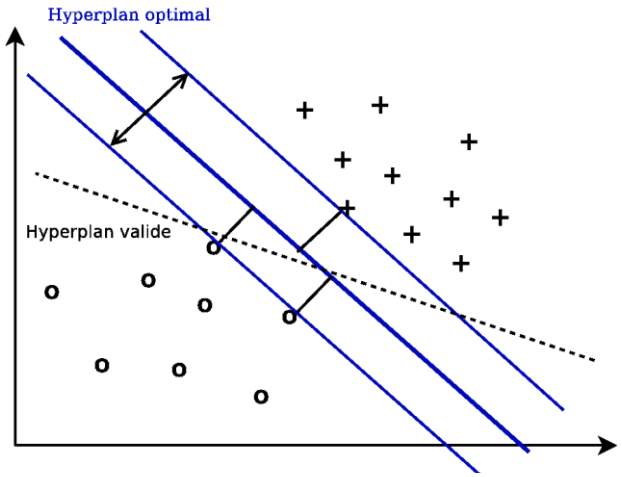


Figure III-6 Hyperplan séparateur à marges maximales

Dans la plupart des problèmes rencontrés en pratique, les données ne sont cependant pas séparables par un hyperplan. Il convient alors de modifier l'approche précédente afin de prendre en compte la possibilité d'observations mal classées. Pour permettre aux SVM de répondre à ces problèmes complexes de classification, les algorithmes initiaux ont été transformés pour élaborer des structures de détection non-linéaires. L'extension au cas non-linéaire peut être effectuée en transformant les observations à l'aide d'une fonction Φ puis en appliquant un détecteur linéaire. La fonction Φ est implicitement définie par le choix d'un noyau K tel que $K(x_i, x_i') = \langle \Phi(x_i) ; \Phi(x_i') \rangle$ où x_i est une observation tirée de la base d'apprentissage χ. La fonction K est donc définie ainsi :

$$K : \chi \times \chi \rightarrow \mathbb{R}$$
$$(x_i, x_j) \mapsto K(x_i, x_j)$$

[1] Support Vector Machines, en français Machines à Vecteurs de Support, ou Séparateurs à Vastes Marges.

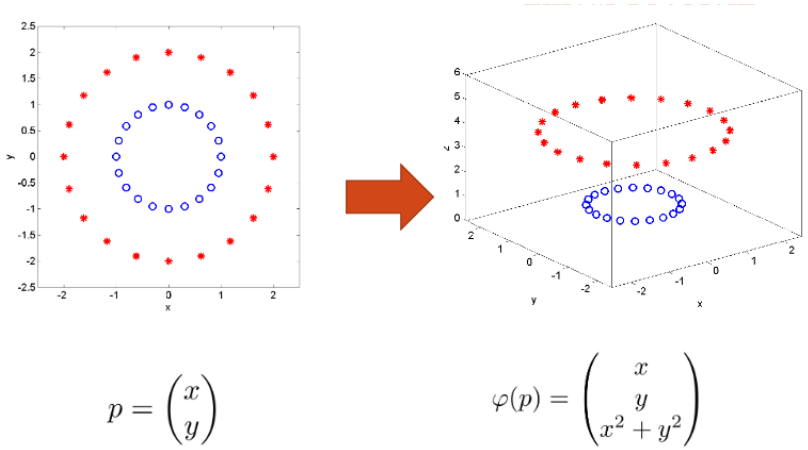


Figure III-7 Illustration du *Kernel Trick* permettant de ramener un cas de séparation non-linéaire (à gauche) à un problème linéaire (à droite)

Les noyaux permettent donc d'étendre aisément au cas non-linéaire des techniques d'apprentissage initialement développées pour le cas linéaire (voir Figure III-7). Le choix du noyau et de ses paramètres se fait généralement d'une manière heuristique lors de tentatives du type essai-erreur. La recherche de méthodes reposant sur des arguments théoriques solides n'en est pas moins un des défis à relever dans l'avenir des méthodes à noyaux. Une liste exhaustive de noyaux reproduisants et des développements supplémentaires sur ces noyaux, comme par exemple la combinaison de noyaux, peut être consultée dans [Souza, 2010; Vapnik, 1995], nous ne présentons ici que quelques-uns des noyaux les plus utilisés.

- **Noyaux polynômiaux** : Les noyaux polynômiaux se traduisent par une règle de décision reposant sur une statistique polynômiale de degré q. Parce qu'ils sont fonction du produit scalaire des observations, de tels noyaux sont dits projectifs. Ils sont exprimés par des fonctions :

$$K(x_1, x_2) = (c + \langle x_1 ; x_2 \rangle)^q$$

- **Noyaux gaussiens** : Les noyaux gaussiens sont des noyaux de type radial, indiquant qu'ils dépendent de la distance $\|x_1 - x_2\|$ entre les observations. Ces noyaux sont définis par :

$$K(x_1, x_2) = \exp\left(-\frac{\|x_1 - x_2\|^2}{\beta}\right)$$ où β est appelé largeur de bande.

- **Noyaux exponentiels** : Le noyau exponentiel est un autre exemple de noyau de type radial. Il est défini par :

$$K(x_1, x_2) = -\|x_1 - x_2\|^q$$

- **Noyaux sigmoïdaux** : Le noyau sigmoïdal dépend de deux paramètres et est défini comme étant :

$$K(x\_1, x\_2) = \tanh(\alpha - \langle x_1; x_2 \rangle + \beta)$$

2.4.2 Autres classifieurs

- **K Nearest Neighbours**[1] : Cet algorithme consiste à mémoriser les exemples de chaque classe dans le corpus d'apprentissage, puis lors de la classification à identifier les k exemples les plus proches de l'instance selon une mesure de distance ou de similarité. Le choix d'une métrique adaptée est essentiel, certaines implémentations utilisent la distance Euclidienne, d'autres la distance de Manhattan ou encore celle de Hamming. L'estimation de la classe est enfin obtenue par un vote sur la valeur qui revient le plus souvent parmi les k voisins. Cette approche est simple et efficace, mais assez coûteuse en raison du calcul, pour chaque instance, de sa distance avec l'ensemble des exemples mémorisés lors de l'apprentissage.

- **Modèles de Markov** : Les modèles de Markov (HHM[2] et MDC[3]) sont des modèles génératifs qui peuvent être vus comme des automates probabilistes à états finis, constitués d'un ensemble d'états et de transitions modélisés par un graphe. Ils sont très populaires dans les domaines du traitement de la parole, en reconnaissance de formes, en bio-informatique et commencent à être de plus en plus utilisés pour la classification d'images et de séquences vidéo [Boreczky and Wilcox, 1998; Eickeler and Müller, 1999; Naphade and Huang, 2000].

- **Réseaux bayésien** : Les réseaux bayésiens sont des modèles probabilistes assignant aux exemples la classe à la probabilité a posteriori maximale. Ils sont des classifieurs simples bien adaptés à des distributions gaussiennes, et très utilisés pour la reconnaissance d'objets. Une étude de ces approches est disponible dans [Jain et al., 2000] et une application de réseaux bayésien naïfs au problème de classification d'images est proposée dans [Naphade and Huang, 2000].

[1] Ou K plus proches voisins.
[2] Hidden Markov Model, ou Modèles de Markov Cachés.
[3] Markov Decision Problem, ou Modèles de Décision Markoviens.

- **Réseaux de neurones** : Un réseau de neurones est un assemblage de neurones artificiels interconnectés. Un réseau réalise une ou plusieurs fonctions algébriques de ses entrées, par composition des fonctions réalisées par chacun des neurones. La capacité de traitement de ce réseau est stockée sous forme de poids d'interconnexions obtenus par un processus d'apprentissage à partir d'un ensemble d'exemples [Philbin et al., 2008]. Ces techniques ont montré de très bonnes performances, notamment pour la détection de visages dans [Rowley et al., 1998].

On observe enfin parfois d'autres méthodes de classification plus marginales. Citons par exemple l'utilisation d'arbres de décision flous (*Fuzzy Decision Tree*) pour la découverte des caractéristiques discriminantes [Breiman, 2001], et la classification d'images [Marsala and Detyniecki, 2005, 2006]. L'équipe d'Oxford a également proposé une méthode d'identification de concepts basée sur les *Random Forest* [Philbin et al., 2008]. Cependant ces approches ne parviennent pas pour l'instant à des résultats comparables aux SVM. D'autres algorithmes tels que les colonies de fourmis ou les essaims de particules ont également été envisagés [Zhang et al., 2008] mais ces voies restent peu développées et souffrent souvent de performances relativement faibles.

2.5 Localisation

La classification d'objets (ou de concepts) se définit comme la catégorisation (positive ou négative) d'une image selon qu'elle contienne ou non une ou plusieurs instances de la classe considérée. Exprimée en langage courant, elle permet par exemple de répondre à la question « y'a-t-il une maison dans l'image ? ». Depuis 2003 les algorithmes de classification ont connu des améliorations considérables, comme le montre l'augmentation des scores au sein de compétitions telles que le Pascal Visual Object Classes Challenge [Everingham et al., 2009]. Une tâche différente, la détection d'objets (ou localisation), consiste non seulement à juger de la présence d'instances de la classe considérée, mais également à déterminer leur position spatiale dans l'image. Pour reprendre l'exemple précédent il s'agirait de fournir la taille et l'emplacement de la ou des maisons au sein de l'image. Cette tâche s'avère beaucoup plus difficile que la première, et les performances de détection n'ont pas suivi la même évolution que celles de classification, restant toujours bien inférieures [Chum and Zisserman, 2007].

Pour résoudre ce problème, une grande majorité des approches reprend l'architecture présentée dans cette section, reposant sur le pipeline de de détection de points d'intérêt, d'extractions de descripteurs locaux autour de ceux-ci puis de classification, en appliquant les classifieurs sur des sous parties de l'image par des méthodes de fenêtre glissante [Everingham et al., 2009]. Celles-ci consistent à découper

l'image en zones de taille fixe, centrées sur chacun des pixels, pour différentes échelles. Certains travaux proposent des heuristiques pour éviter une recherche trop coûteuse à travers un nombre de fenêtres trop important. Il est par exemple possible d'utiliser une pré-segmentation de l'image pour sélectionner des régions candidates [Viitaniemi and Laaksonen, 2006], de prendre en compte la position des points d'intérêt les plus discriminants pour la classe [Chum et al., 2007] , ou encore d'utiliser des méthodes branch-and-bound [Lampert et al., 2008].

3. Spikenet MultiRes, une approche bio-inspirée

Au cours du siècle dernier, les sciences du traitement de l'information et du cerveau ont évolué d'une manière interdépendante. Ainsi, comme le soulignait le mathématicien Von Neumann, à l'origine de la célèbre architecture du même nom utilisée dans la quasi-totalité des ordinateurs modernes, le cerveau est traditionnellement vu et étudié comme un système de traitement de l'information -les neurones étant considérés comme ses unités élémentaires de calcul-, alors qu'en parallèle de multiples aspects de l'informatique et des mathématiques se sont inspirés des connaissances sur le fonctionnement et l'organisation du système nerveux [Neumann, 1958]. Malgré l'augmentation exponentielle de la puissance de calcul des processeurs, les systèmes biologiques continuent pourtant de surpasser les traitements informatiques dans de très nombreuses tâches telles que l'interprétation des scènes visuelles et du langage, ou encore le contrôle moteur [Mussa-Ivaldi and Miller, 2003]. Ces performances exceptionnelles du vivant expliquent l'intérêt croissant pour les méthodes bio-inspirées depuis une cinquantaine d'années dans de nombreux domaines. Celles-ci, en modélisant des systèmes artificiels reproduisant certains aspects de l'architecture et des traitements du cerveau, espèrent ainsi se rapprocher de ses performances computationnelles [Cox and Pinto, 2011].

Au niveau de la vision, les recherches de David Marr ou de Tomaso Poggio, qui comptent parmi les pionniers de cette branche des neurosciences computationnelles [Marr, 1982; Marr and Hildreth, 1980; Marr and Poggio, 1976; Poggio and Edelman, 1990; Serre et al., 2005], ont initié une longue tradition d'algorithmes de vision et d'apprentissage automatique bio-inspirés.

Dans le domaine de la reconnaissance d'objets et de la classification d'images, ce sont les réseaux de neurones convolutionnels (voir Figure III-8) qui se sont imposés comme les solutions les plus utilisées [Ciresan et al., 2012; LeCun et al., 2010, 2004; Le et al., 2011; Riesenhuber and Poggio, 2002; Rowley et al., 1998]. Comme nous le détaillons en annexe de cette thèse, les premiers traitements du système visuel sont de nature hiérarchique. Les cellules de la rétine projetant vers les neurones des corps géniculés latéraux, qui propagent ensuite l'information visuelle vers le cortex visuel primaire où la combinaison de leurs activations produit des cellules sélectives aux orientations : les cellules simples. Celles-ci permettent à leur tour d'activer des cellules complexes, qui serviront d'entrée aux cellules hypercomplexes (ou *end-stopped*) [Würtz and Lourens, 2000, 1997]. La succession de traitements en remontant la voie ventrale permet d'acquérir des propriétés d'invariance et une complexification croissante des champs récepteurs aboutissant dans le cortex inféro-

temporal à des neurones codant des stimuli haut niveau tels que des objets ou des visages. Pourtant, Les opérations de base de ces mécanismes sont de simples sommes de sorties et non-linéarités, qui peuvent être facilement modélisées dans un réseau de neurones *feedforward*. Ceci explique la tendance actuelle à multiplier les couches de traitement comme on peut l'observer dans les méthodes de Deep Learning.

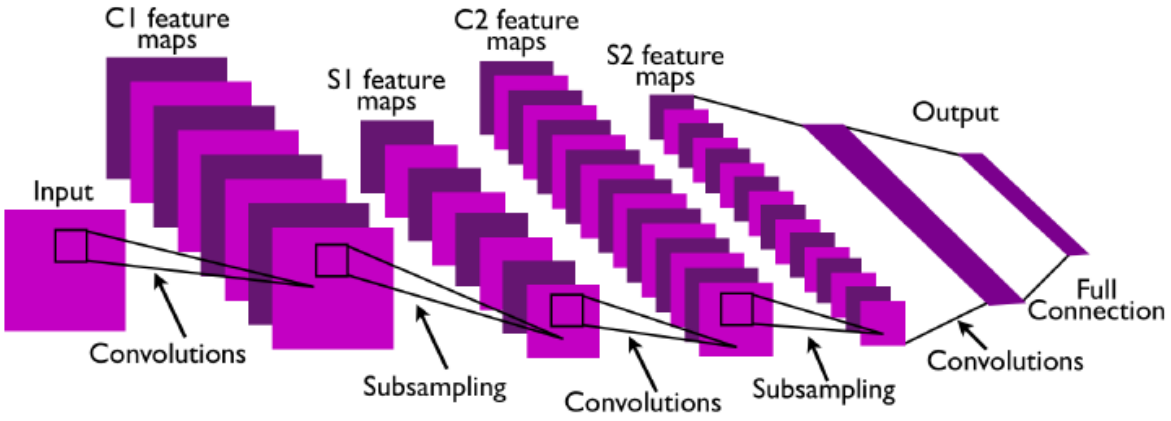


Figure III-8 Architecture standard d'un réseau de neuronnes convolutionnels [LeCun et al., 2010]

A l'inverse de cette approche, l'algorithme Spikenet s'inspire des mécanismes de reconnaissance ultra-rapides réalisés par les couches initiales du système visuel. Développé au Centre de Recherche Cerveau et Cognition, il reproduit ces premières couches de traitement du système visuel à l'aide de neurones impulsionnels et asynchrones. Simon Thorpe, Ruffin Van Rullen et Arnaud Delorme ont mis au point ce modèle vers la fin des années 90 en se basant sur des données concernant les performances du système visuel dans différentes tâches, sur les propriétés de neurones observés à différents niveau de la voie ventrale, sur la connectivité anatomique dans différentes aires corticales et sur les aspects biophysiques au niveau cellulaire. Si celui-ci se voulait à l'origine un outil de simulation pour la validation d'hypothèses sur le fonctionnement cérébral, ses performances surprenantes ont mené à la création d'une entreprise et à l'industrialisation de cette technologie, amenant de nombreuses améliorations et optimisations de l'algorithme originel.

L'architecture de Spikenet, décrite dans [Delorme et al., 1999; Delorme and Thorpe, 2003; Thorpe et al., 2004; VanRullen et al., 1998], repose sur le constat que les très faibles latences observées dans des tâches de catégorisation visuelle ne sont compatibles qu'avec l'envoi des premiers potentiels d'action (ou spikes) à chaque étape neuronale. Cette question avait été soulevée pour la première fois dans [Thorpe and Imbert, 1989]. De nombreuses études ont montré depuis des latences extrêmement faibles, de l'ordre d'une centaine de millisecondes, aussi bien au niveau comportemental que physiologique, pour diverses tâches

visuelles impliquant la détection ou la catégorisation de visages, d'animaux, de véhicules ou encore de scènes naturelles. Il en découle, d'une part, que l'information transmise se base sur l'activation ou non des neurones plus que sur leurs valeurs de sortie transmises (se manifestant par un taux de décharge moyen ou une latence précise), et d'autre part que comme la latence de décharge d'un neurone est inversement fonction de son niveau d'activation, seuls les neurones avec l'activation la plus haute généreront des potentiels d'action assez tôt pour être pris en compte par le relais suivant [VanRullen and Thorpe, 2001]. A l'inverse de la plupart des autres modèles utilisant un codage par fréquence de décharge, Spikenet repose donc sur un codage par rang, détaillé dans [Thorpe et al., 2001; VanRullen et al., 2005; VanRullen and Thorpe, 2002]

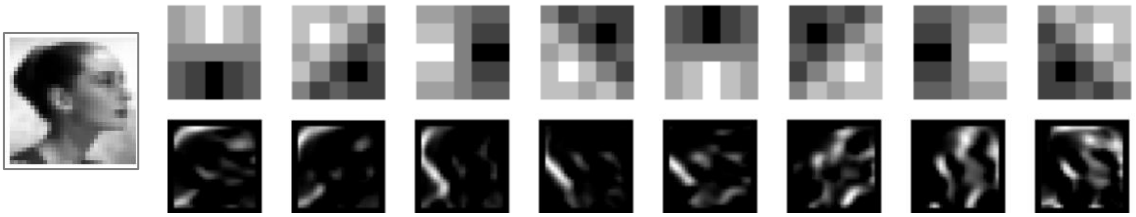


Figure III-9 Masques de convolutions (1ere ligne) et cartes d'activations correspondantes (2eme ligne) pour l'image représentée à gauche

Ces propriétés ont été intégrées à l'algorithme Spikenet et se traduisent du point de vue de l'implémentation par un modèle reproduisant les traitements de l'aire V1, dont seuls les premières activations seront conservées et transmises à la couche supérieure de classification. La zone à apprendre est d'abord redimensionnée en une vignette de 900 pixels, puis filtrée par une banque de 8 convolutions (4 orientations et 2 polarités, cf. Figure III-9), similaires aux champs récepteurs trouvés dans le cortex visuel primaire, codant l'information visuelle sous forme de lignes et d'arêtes de différentes orientations [Atick and Redlich, 1990; Daugman, 1985; Jones and Palmer, 1987]. Les 255 valeurs les plus hautes[1] parmi ces 8 cartes de convolutions sont ensuite sélectionnées de façon itérative, et conservées au sein d'une structure de données. La méthode de sélection utilisée s'apparente à la poursuite gourmande proposée par Laurent Perrinet pour l'encodage d'images basé sur la latence selon des modèles de rétine ou de V1 [Laurent Perrinet et al., 2004; L. Perrinet et al., 2004; Perrinet, 2004]. Inspirée par l'algorithme statistique de *Matching Pursuit* [Mallat and Zhang, 1993], elle consiste à sélectionner le neurone à la réponse la plus forte, appliquer les inhibitions latérales résultant de cette activation, puis finalement soustraire celle-ci et rechercher la nouvelle valeur maximale. Ce processus récursif est appliqué dans le noyau Spikenet en respectant les règles d'apprentissage suivantes :

[1] Correspondant par conséquent aux neurones ayant déchargé les plus tôt.

- Compétition locale : en chaque pixel seule l'orientation la plus forte est conservée (méthode *winner take all* [Coultrip et al., 1992])
- Inhibition globale de la population de neurones similaires à celui ayant déchargé (observée notamment lors d'enregistrement unitaires de cellules de V1 chez le singe [Knierim and Van Essen, 1992]), assurant ainsi une répartition relativement homogène des différentes orientations dans le modèle final ;
- Inhibition locale des neurones adjacents, quelle que soit leur sélectivité, afin de répartir spatialement l'activité dans l'ensemble de l'image à apprendre (et par exemple éviter une trop grande concentration dans une sous-partie de l'image particulièrement contrastée) ;

Le modèle[1] résultant de cet apprentissage, illustré dans la Figure III-10, ne conserve que l'orientation dominante en chacun des 255 points sélectionnés, et non les valeurs de convolution ou leur ordre de décharge. Il perd donc l'information sur le niveau d'activation des neurones, pour ne retenir que les saillances principales constituant le motif appris.

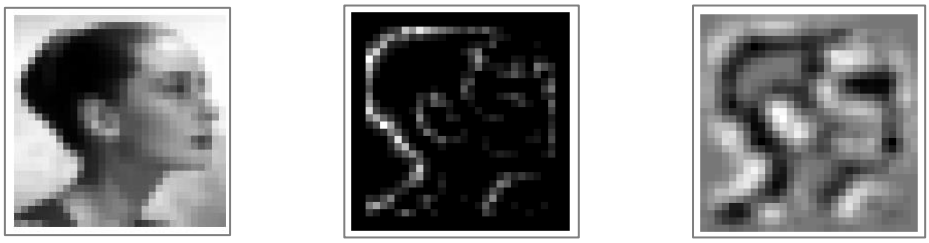


Figure III-10 De gauche à droite, image originale, carte d'activation avec les 255 poids les plus haut, toutes orientations confondues, et enfin reconstruction de l'image à partir du modèle créé

Lors de la recherche d'un modèle, l'image testée est elle aussi filtrée selon les 8 mêmes masques de convolution, en ne conservant pour chacun des pixels que l'orientation dominante. Un système de vote des différents neuronnes du modèle en chacun des points de l'image permet ensuite de déterminer sa probablité de détection aux différentes positions possibles, ces scores correspondant au pourcentage de neurones du modèle ayant été activés. Si dans cette phase de reconnaissance un modèle peut être recherché à différentes échelles en rééchantillonant l'image testée, l'information visuelle encodée correspond malgré tout à une seule résolution, celle utilisée lors de l'apprentissage (30 par 30 pixels), et ne couvre donc qu'une faible plage de fréquences spatiales.

[1] Dans la terminologie de l'algorithme Spikenet, on entend par modèle l'activité caractéristique d'un motif visuel appris, c'est-à-dire le schéma de décharge des neurones les plus rapides, stocké dans une structure de données. Nous nous réfèrerons donc souvent à cette définition par la suite, plutôt qu'au sens traditionnel de modèle de traitement, ou modèle computationnel.

Pour tenter d'améliorer les performances de cet algorithme, nous allons développer dans cette section la mise en place d'une nouvelle architecture baptisée Spikenet MultiRes, combinant l'information spatiale disponible à différentes résolution, dans un modèle en cascade *coarse to fine.*

3.1 Etude préliminaire sur l'architecture MultiRes

Le système visuel humain est capable de détecter et de localiser certains stimuli en vision périphérique à de très faibles latences et avec une grande précision, permettant ainsi de guider les saccades oculaires vers ces zones d'intérêt pour un traitement ultérieur plus poussé en vision centrale. Ces résultats suggèrent des mécanismes relativement simples et rapides impliquant un nombre restreint d'étapes de traitement, probablement purement feedforward pour les réponses comportementales les plus rapides, et reposant sur une quantité faible d'informations étant donnée la taille réduite des stimuli à des excentricités élevées où l'acuité visuelle et la taille des champs récepteurs limitent fortement la résolution spatiale du stimulus.

Pour intégrer ces constats à l'algorithme de reconnaissance de formes Spikenet dont nous venons de présenter le fonctionnement, nous proposons donc la mise en place d'une nouvelle architecture combinant plusieurs passes de traitement à des résolutions différentes, débutant par une passe certes moins précise mais très rapide à une échelle faible, telle que semble être effectuée la détection de stimuli en vision périphérique. Nous espérons ainsi pouvoir d'une part augmenter la vitesse de traitement en limitant le nombre de modèles testés à des résolutions « importantes » et en contraignant l'espace de recherche du champ visuel aux zones candidates sélectionnées par les résolutions basses ; d'autre part augmenter la fiabilité des détections en renforçant la quantité de signal sur les motifs recherchés par l'information supplémentaire fournie à ces différentes fréquences spatiales ; et enfin améliorer la robustesse des détection grâce à la tolérance plus importante des basses résolutions à de nombreuses déformations.

3.1.1 Choix des résolutions

L'architecture classique Spikenet repose, comme nous l'avons dit, sur des patchs d'images de 900 points (ou pixels). Par souci de clarté nous nous intéresserons par la suite à des modèles carrés, qui pour cette configuration ramèneraient donc les patchs d'images à une dimension de 30 par 30 points. Le cas de rectangles ou de figures plus complexes par le biais de masques n'ayant pas d'incidence sur la nature et les performances des traitements,

il est plus simple de se limiter dans cette étude à ce format carré, pour une lecture des résultats plus aisée.

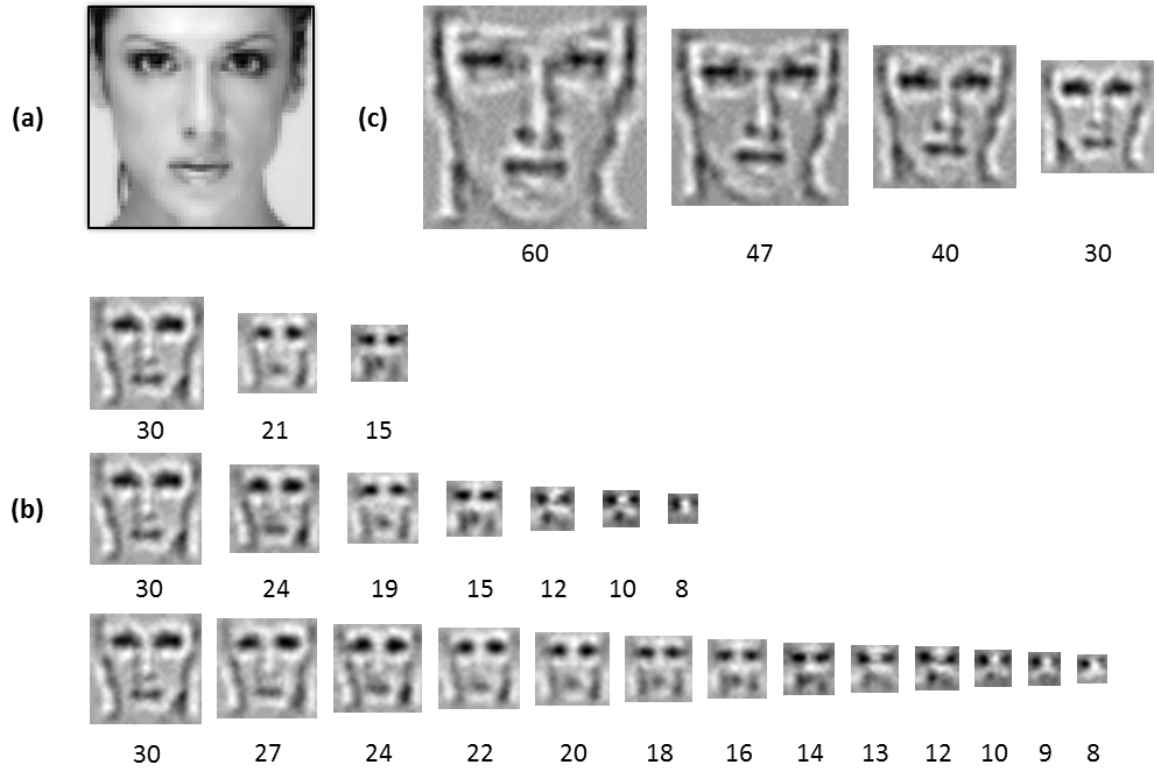


Figure III-11 (a) Image originale ; (b) et (c) différents choix de taille et de nombre de résolutions (les images présentées sont le résultat de reconstructions à partir des modèles Spikenet appris)

Parmi les nombreuses possibilités relatives à cette nouvelle architecture multi-résolutions, une des premières questions à poser est évidemment le choix du nombre et de la taille de ces résolutions. Jusqu'à quel point pouvons-nous réduire la résolution des échelles les plus basses afin de conserver suffisamment d'information et ne pas détériorer les performances ? Combien d'échelles sont nécessaires ? Si en augmentant leur nombre on augmente la quantité de signal relatif à la cible à apprendre, l'information extraite étant différente à chacune des fréquences spatiales correspondantes, un nombre trop important aura pour conséquence non seulement une redondance d'information, mais également un coût en termes de temps de traitement. Enfin si le choix de 30 par 30 pixels dans le noyau Spikenet traditionnel est le résultat d'évaluations ayant montré qu'il s'agissait d'un compromis optimal entre vitesse d'exécution et pouvoir discriminant, peut-on imaginer dans cette nouvelle architecture l'utilisation de résolutions plus importantes afin d'obtenir un niveau de détail plus conséquent sans trop de pénalités en termes de coûts computationnels ? La Figure III-11 illustre ces différentes possibilités en présentant la reconstruction de modèles appris à partir de l'image de gauche à différentes résolutions.

3.1.2 Répartition des poids entre échelles

En plus de déterminer le nombre et la taille des différentes échelles, un autre facteur à prendre en considération est le choix du nombre de spikes (ou poids) qui seront conservés. Dans le noyau classique, parmi les 900 valeurs résultant de l'apprentissage (pour rappel, chacun de ces 900 poids correspond à l'orientation dominante en un pixel), seuls 255 sont retenus, ceux dont les valeurs de convolution sont les plus hautes, ce qui revient dans un schéma de codage par rang à ne conserver que les premiers spikes, ceux des neurones les plus activés, ayant donc déchargé le plus tôt. La valeur retenue est un résultat empirique fixé lors du développement du noyau Spikenet qui a montré de bonnes performances en détection et autorisant certaines optimisations matérielles grâce au stockage des valeurs sur un octet.

Pour illustrer la problématique du nombre de poids dans le cas d'une architecture multi-résolutions prenons à titre illustratif une configuration combinant 7 différentes résolutions, ici <8;10;12;15;19;24;30>. En respectant le même pourcentage de poids conservés pour chaque échelle que celui du noyau classique, nous obtiendrons les valeurs reportées dans la colonne *Poids_1* de le Tableau III-1, soit un nombre total de spikes de 582. La colonne suivante, *Poids_2* correspond à une normalisation de leur somme, tout en respectant les mêmes pourcentages à chaque échelle, afin de contraindre ce nombre total de poids par modèle à 256.

Taille	Points	Poids_1	Poids_2
30	900	256	113
24	576	151	66
19	361	85	37
15	225	46	20
12	144	24	11
10	100	14	6
8	64	6	3
total		*582*	*256*

Tableau III-1 Nombre de poids par échelle

Une autre alternative, plus biologiquement plausible et se rapprochant de certains mécanismes du noyau Spikenet serait une compétition inter-échelles. Ainsi pour un nombre total de poids *N* donné, seuls les *N* premiers neurones à décharger seraient conservés, peu importe leur résolution d'origine. Dans un tel schéma, le nombre de spikes pour chaque résolution serait donc différent pour chaque modèle. Différentes simulations nous ont

permis de conclure que la répartition des poids avec une stratégie de compétition s'avère très proche d'un nombre fixe par échelle respectant le même pourcentage que celui mentionné précédemment. La Figure III-12 Illustre ces résultats pour 4 images de test.

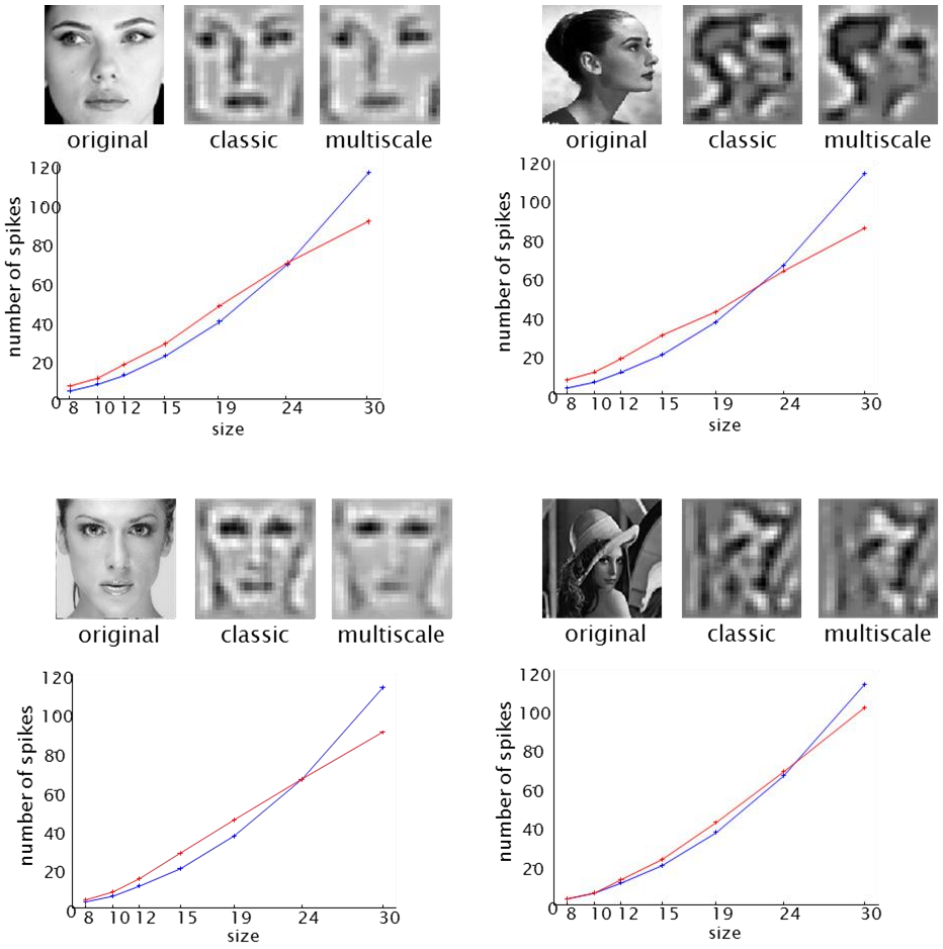


Figure III-12 Nombre de poids sélectionnés pour chaque échelle lors de l'apprentissage du modèle à chaque échelle. En bleu lorsque ce nombre est fixé proportionnellement à la résolution, en rouge suite à une compétition inter-échelle; les images au-dessus correspondent dans l'ordre à l'image apprise, la reconstruction à partir du noyau classique, et celle à partir du noyau MutltiRes

Afin d'observer le comportement de ces différents choix d'apprentissage et d'en tirer certaines observations, étant encore dans une phase initiale de mise en place de l'architecture, nous nous sommes intéressés aux résultats des reconstructions à partir des modèles appris. Nous avons donc pour chaque échelle un certain nombre de poids retenus constituant le modèle. Ceux-ci correspondent à l'orientation dominante pour un pixel donné. La reconstruction consiste donc à ajouter pour chacun de ceux-ci le masque correspondant à la convolution de l'orientation sélectionnée centré aux coordonnées du neurone. Les images résultantes pour chacune des résolutions sont finalement combinées entre elles par une

simple somme pondérée. Les résultats de ces reconstructions pour les différents choix possibles de poids sont présentés dans la Figure III-13.

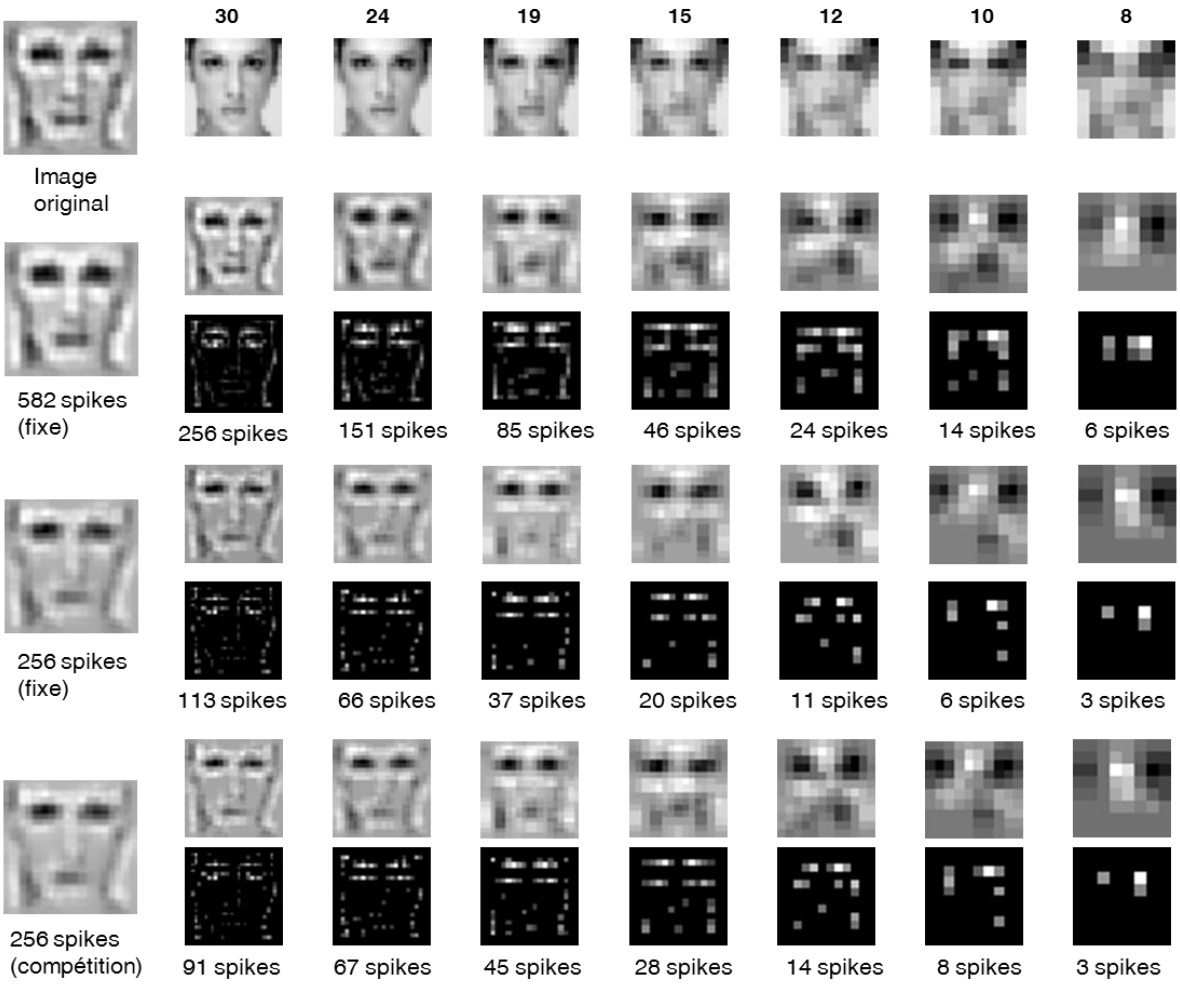


Figure III-13 Reconstruction des modèles. L'image à gauche correspond à la reconstruction finale (combinant les différentes échelles), les premières lignes aux reconstructions à chaque échelle, et les deuxièmes aux cartes d'activation pour chacune (les poids retenus dans le modèle sont indiqués en couleur claire)

Comme nous pouvons le voir dans les cartes d'activation comme dans les reconstructions par échelle, chaque résolution extrait des informations différentes sur la cible, correspondant à des fréquences spatiales spécifiques. En allouant un nombre suffisant de poids le modèle résultant est suffisant dense et semble, du moins visuellement, représenter de manière satisfaisante le stimulus d'entrée. En revanche si, comme dans l'architecture classique, on contraint le nombre total de poids à 256 (avec ou sans compétition entre échelles, car comme nous l'avons vu précédemment leur résultats sont très proches), on observe alors une perte d'information. Les poids étant répartis à travers les échelles, leur nombre pour chacune s'avère alors trop faible, à plus forte raison si

l'information est redondante entre les échelles. Ce phénomène est visible dans la Figure III-13 mais encore plus notable dans la Figure III-14. On y remarque que les cartes d'activation de résolutions voisines sont effectivement très similaires. Pour pallier ce constat deux solutions sont envisageables. Nous pouvons augmenter le nombre total de poids, de sorte à avoir un nombre suffisant de spikes pour chaque échelle, ou bien mettre en place des stratégies d'inhibition entre échelles, afin de limiter cette redondance.

Pour explorer cette piste d'inhibition entre résolutions nous avons procédé à un certain nombre de tests en jouant sur différents paramètres. Soit procéder à une suppression complète des neurones à la même position spatiale, comme dans la compétition entre orientations pour un point donné, soit simplement réduire leur activité, tel que cela est fait pour l'inhibition locale. Un autre facteur pouvant influencer l'apprentissage serait que cette inhibition partielle ou complète s'applique seulement à l'orientation considérée ou à toutes. Imaginons qu'en un point de l'espace une arête verticale soit détectée à faible résolution, pour la même position à des plus hautes fréquences spatiales : souhaite-t-on éviter tout autre spike pour assurer une répartition plus homogène dans l'ensemble du champ récepteur, ou seulement les spikes correspondant à des orientations verticales afin de limiter la redondance entre échelles? Un dernier paramètre qu'il convient de fixer est la « portée » de cette inhibition. La similitude entre les activations à différentes résolutions est de façon évidente beaucoup plus importante pour des échelles voisines. Il semble donc plus judicieux, au lieu de répercuter cette inhibition sur l'ensemble des autres échelles, de se limiter aux échelles voisines. Dans le souci de limiter le nombre total de poids, nous avons constaté qu'une solution optimale consistait à inhiber seulement les échelles voisines de plus hautes résolutions, afin de privilégier les basses fréquences, moins coûteuses en nombre de spikes.

L'évaluation de la qualité de ces différents essais étant empirique, en observant les différentes cartes d'activation et leur reconstruction, il n'est pas possible d'affirmer de manière absolument fiable l'optimalité de ces combinaisons de paramètres. Néanmoins les choix semblant donner les meilleurs résultats sont une inhibition partielle des neurones des deux échelles supérieures aux mêmes positions, et ce toutes orientations confondues. La Figure III-14 présente ces résultats. Il apparaît en conclusion que, même sans inhibition inter-échelles, l'on obtienne une reconstruction fidèle du stimulus pour un nombre total de spikes suffisamment élevé. En revanche pour un nombre limité, la redondance entre échelles voisines est flagrante, avec pour résultat une importante perte d'informations aux hautes fréquences spatiales, qui se matérialise ici par l'absence de détail au centre du visage. En revanche, avec la mise en place de mécanismes d'inhibition inter-échelles, si le nombre de spikes par résolution reste assez proche, ceux-ci se répartissent différemment, et l'on parvient ainsi à capter les informations importantes aux résolutions les plus grandes.

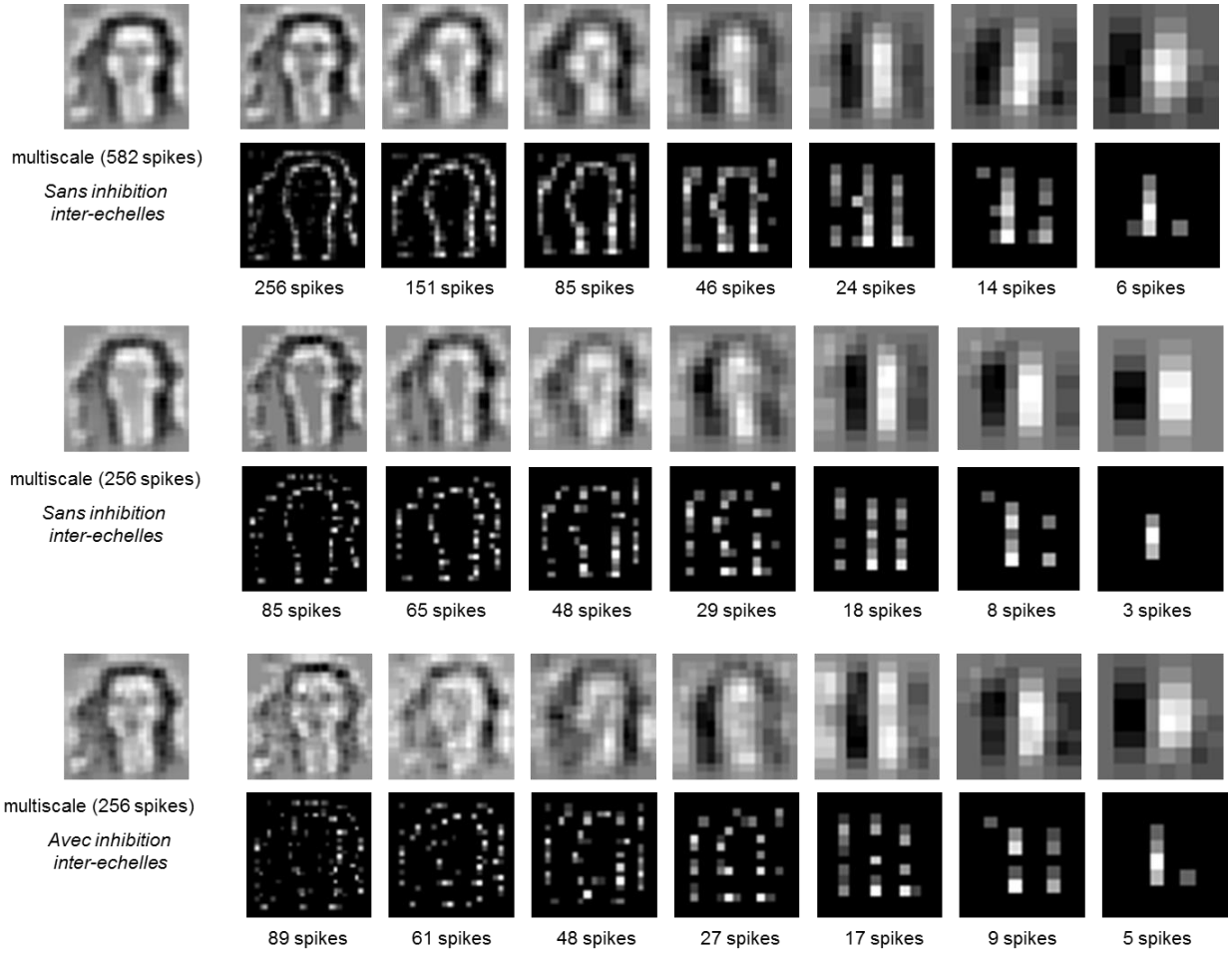


Figure III-14 Reconstruction des modèles. L'image à gauche correspond à la reconstruction finale (combinant les différentes échelles), les premières lignes aux reconstructions à chaque échelle, et les deuxièmes aux cartes d'activation pour chacune (les poids retenus dans le modèle)

3.2 Méthodes

L'évaluation et la comparaison d'algorithmes de vision par ordinateur reste un sujet délicat dans la communauté scientifique, posant de nombreux problèmes de méthodologie. Tout d'abord parce qu'ils sont souvent orientés pour des contextes d'application particuliers. Rares sont les méthodes génériques pouvant s'appliquer à n'importe quelle situation, et il n'est donc naturellement pas cohérent de tenter de comparer des solutions à visées différentes. Deuxièmement car, comme souligné dans [Heath et al., 1997], leurs performances dépendent de nombreux facteurs :

1. l'algorithme en tant que tel,
2. les images utilisées pour les tests,
3. les paramètres de l'algorithme choisi,
4. la méthode et les métriques d'évaluation.

Cette évaluation des performances pose aussi la question des qualités attendues d'un algorithme de vision, qui une nouvelle fois dépendent fortement du domaine d'application, et de leur caractérisation par des indicateurs quantitatifs pouvant être calculés et interprétés. Une liste des propriétés généralement recherchées est proposée dans [Wirth et al., 2006] :

1. *précision* : la correspondance entre les prédictions et les résultats attendus ;
2. *robustesse* : la tolérance aux changements ;
3. *sensibilité* : la capacité à discriminer des éléments relativement proches ;
4. *adaptabilité et généralisabilité* : le comportement de l'algorithme face à une variabilité dans les images ;
5. *fiabilité* : la reproductibilité des résultats face aux mêmes données d'entrée ;
6. *efficience* : le coût en terme computationnel (temps de traitement).

Les corpus d'apprentissage ayant une incidence évidente sur les résultats de classification (problème illustré dans la Figure III-15, où la différence de difficulté apparaît clairement en fonction du type d'image utilisé), il convient de comparer les résultats de différents algorithmes sur les mêmes bases d'images. Dans ce but, de nombreuses campagnes d'évaluation ont vu le jour depuis une dizaine d'années, permettant de constituer des collections d'images annotées, certaines même segmentées, de fournir des outils d'évaluation, et une méthodologie standardisée pour stimuler et comparer les algorithmes de vision les plus performants.

Dans cette partie nous détaillerons donc la méthodologie employée pour le développement et l'évaluation de l'algorithme Spikenet MutltiRes, inspirée par les pratiques et métriques répandues dans la communauté.

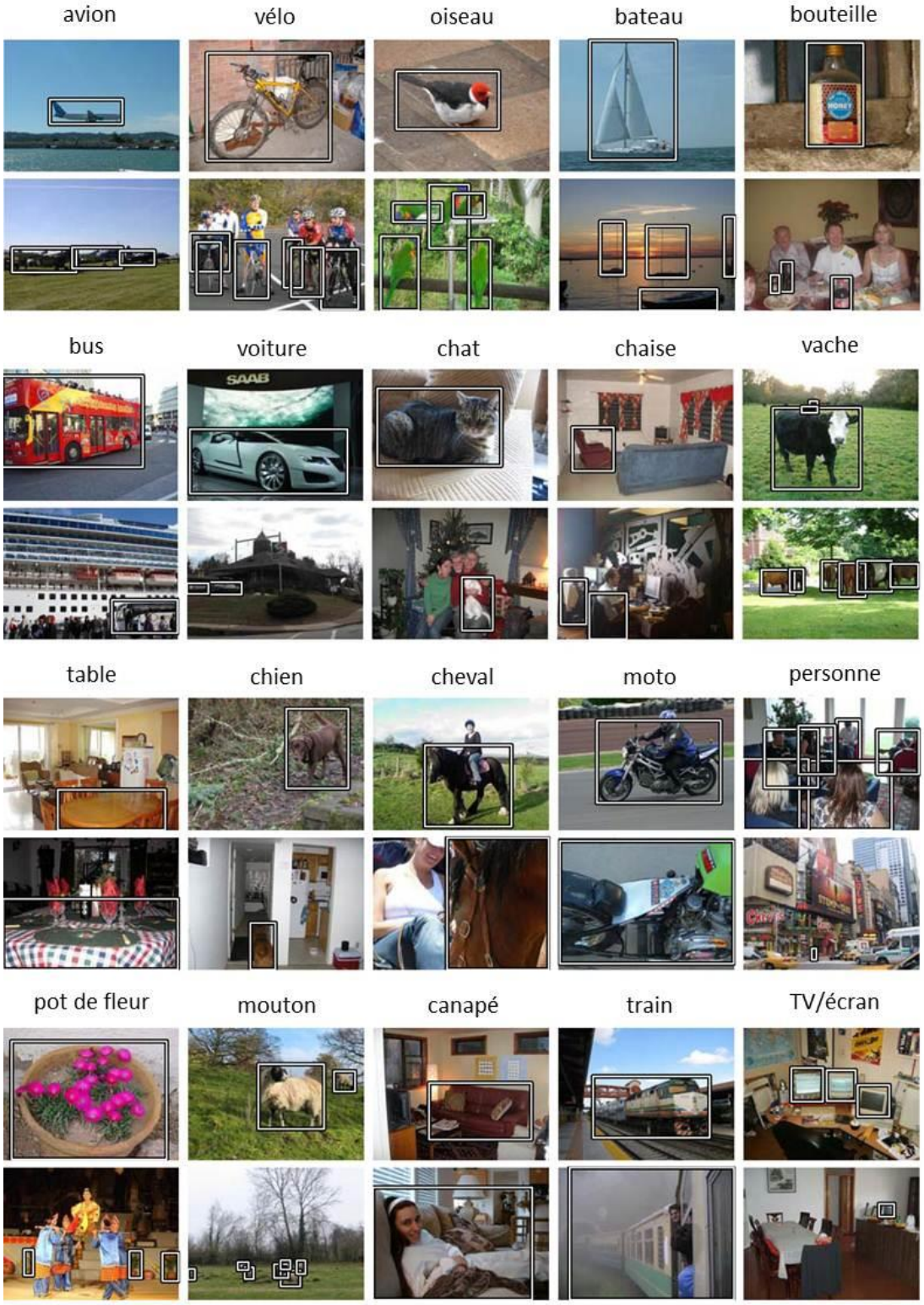


Figure III-15 Images du corpus VOC2007. Pour chacune des 20 classes annotées, 2 images sont présentées, la supérieure étant considérée comme prototypique, celle du dessous comme difficile (en raison d'occlusions, de l'illumination, de la taille, de la qualité de l'image ou de l'angle de vue par exemple). Les rectangles blancs correspondent aux boîtes englobantes (*bounding box)* de chaque objet.

3.2.1 Corpus d'apprentissage et de test

Le corpus d'apprentissage que nous avons constitué comporte 250 images. Il se décompose en 2 sous-ensembles, tout d'abord un set de cent images « complexes » : des images riches et variées contenant aussi bien des scènes d'intérieur, d'extérieur, naturelles ou manufacturées. Celui-ci se veut donc générique, à l'inverse du second, comprenant 150 images de scènes de rues, présentant des immeubles, bâtiments, façades, et autres éléments urbains. Ce choix correspond au cas d'utilisation le plus fréquent au sein du projet Navig présenté dans le deuxième chapitre de ce manuscrit, utilisant la vision artificielle pour l'aide à la navigation des non-voyants.

Chacune de ces images a fait l'objet d'une normalisation de la taille, du contraste et de la luminance afin d'homogénéiser le corpus et d'éviter tout biais lié à ces différents facteurs (voir Figure III-16). La procédure consiste à redimensionner et recadrer les images, en respectant leur aspect-ratio, pour les ramener à une taille de 600 par 600 pixels, puis de les convertir en niveaux de gris, et de fixer la luminance moyenne à 128 ainsi que le contraste RMS[1] à 40. Afin de pouvoir tester un plus grand nombre de positions le motif à apprendre ne correspond pas à l'ensemble de l'image, mais seulement à sa partie centrale. Une vignette de 120 par 120 pixels est donc finalement extraite de celle-ci au centre de chacune.

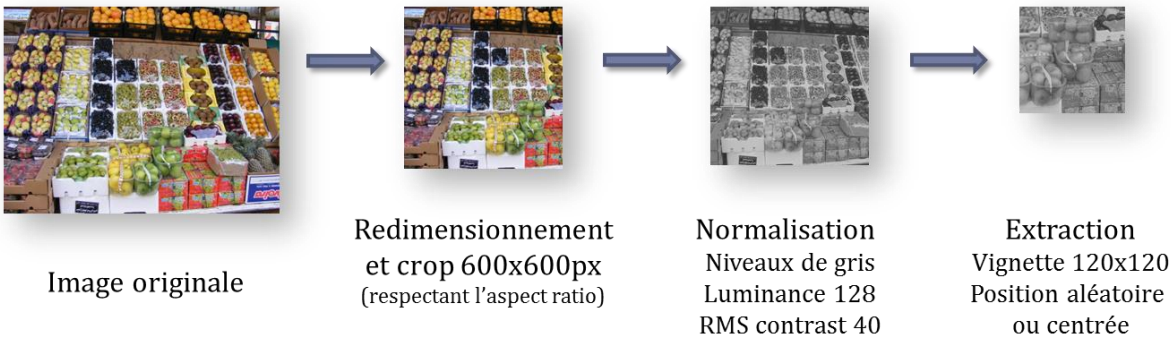


Figure III-16 Procédure de normalisation des images

La base de test est quant à elle composée de ces mêmes 250 images, auxquelles nous avons ajouté 350 images extraites de façon aléatoire parmi une dizaine de films afin d'avoir un plus large éventail de test. Nous avons également complété ce corpus par 250 images de scènes de rue supplémentaires, différentes des 150 mentionnées précédemment, afin d'éprouver le système par l'ajout de distracteurs potentiellement similaires visuellement

[1] Le contraste *Root Mean Square* est définit comme l'écart type des valeurs d'intensité de l'ensemble des pixels de l'image.

aux cibles recherchées. Ces images ont également été normalisées de la même façon que celles du corpus d'apprentissage. En conclusion nous disposons donc de 250 images à apprendre et de 850 à tester. Un échantillon de celles-ci est présenté en annexe.

L'algorithme Spikenet vise à détecter des cibles « uniques » et non des catégories. L'apprentissage est par conséquent effectué sur une image et non un ensemble représentatif d'une classe particulière, comme sont généralement utilisées beaucoup des méthodes classiques de vision artificielle. Celles-ci reposent généralement sur des algorithmes de classification qui extraient les similarités parmi les exemples illustrant un concept donné. Il s'agit du type de tâche le plus fréquent proposé dans les compétitions telles que le Pascal Challenge, TrecVid, ILSVRC, etc. Dans notre cas précis, cela consisterait par exemple à apprendre la catégorie « scène de rue » grâce aux 150 images du corpus d'apprentissage, puis à tenter de différencier dans la base de test de nouvelles images correspondant aussi à des scènes de rue parmi d'autres distracteurs. Ici notre but est différent, nous cherchons à apprendre une scène de rue particulière, identifier un bâtiment par exemple, et à détecter spécifiquement celui-ci, sans réponse sur d'autres qui seraient similaires. Pour évaluer les performances de notre algorithme nous ne disposons donc pour chacune des 250 images de la base d'apprentissage que d'une seule image contenant la cible dans le corpus de test, et de 849 distracteurs (images où la cible n'est pas présente). Nous avons donc besoin d'un nombre plus important d'exemplaires positifs pour chacune, qui permettront de calculer différentes mesures indicatives des performances telles que la précision ou le rappel en fonction des seuils choisis. Pour ce faire nous avons appliqué à chaque image du corpus d'apprentissage différentes combinaisons de transformations. La plupart d'entre elles reflètent les changements d'apparence d'un objet dans des conditions de vue naturelles. Elles peuvent être regroupées en deux catégories, les transformations géométriques et photométriques [Field and Olmos, 2007]. Les premières correspondent à des changements dans la position des pixels dans l'image (lors de translations, rotations, déformations,…), alors que les secondes sont des traitements appliqués aux valeurs de chacun des pixels de l'image (luminance, contraste, bruit,…). L'ensemble des transformations que nous avons appliquées (dont les paramètres ont été tirés aléatoirement), sont présentées ci-dessous et dans la Figure III-17 :

- **Taille** : redimensionnement bilinéaire de l'image (de 50 à 200% de la taille originale).
- **Aspect Ratio** (ou rapport de cadre) : Changement du rapport hauteur/largeur de l'image (de 50% à 150%).
- **Rotation** : rotation de -50° à +50° conservant l'intégralité de l'image originale (pour ces images un masque est aussi généré de sorte que la détection ne soit effectuée que sur la partie correspondant à l'image d'entrée)

- **Perspective** : projection de l'image simulant une vue en perspective. ici deux facteurs entrent en compte, l'angle de perspective, et la magnitude, correspondant à la position du point de fuite (comme pour les rotations des masques ont également été générés pour exclure des traitements le fond gris résultant).
- **Bruit** : pourcentage des pixels dont l'intensité est remplacée par une valeur aléatoire.
- **Flou** : moyennage des pixels par une fenêtre gaussienne glissante (dont la taille peut varier de 1 à 1/20 de la taille de l'image).
- **Contraste** : modification du contraste RMS (de 0 à 100).
- **Luminance** : modification de l'intensité de l'ensemble des pixels (de -100 à +100).
- **Gamma** : valeur de gamma allant de 0 à 2.5 (correspond à l'exposant associée à l'intensité de chacun des pixels).

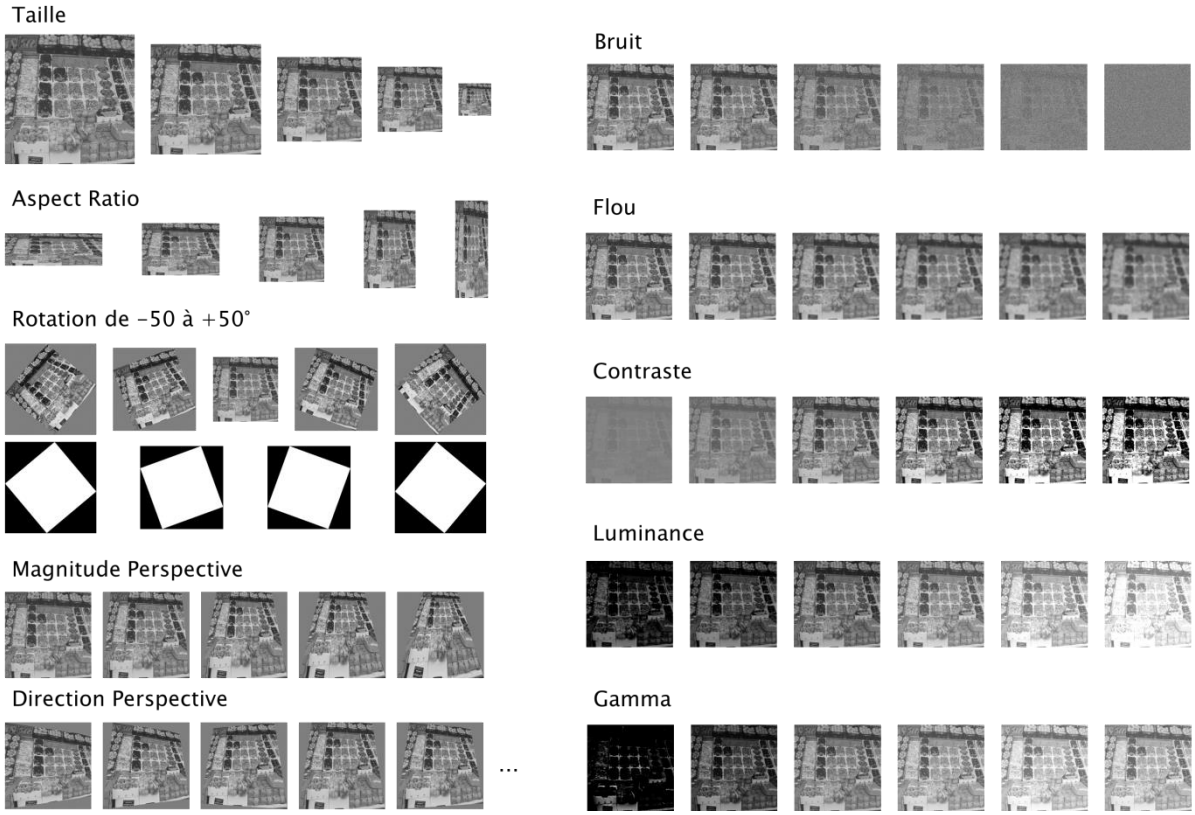


Figure III-17 Illustration des différentes transformations appliquées aux images de la base d'apprentissage

En conclusion, telle que présenté dans la Figure III-18, nous avons donc pour chacune des 250 images du corpus d'apprentissage (d'où ont été extraites les vignettes des motifs à apprendre), 150 images similaires de la même cible, résultat de ces combinaison de traitements. Notons que si ces 150 transformation correspondent à des tirages aléatoires de paramètres pour chacune des modifications présentées (taille, aspect ratio, flou, etc), ces paramètres, une fois déterminés, seront appliqués à chacune des images du corpus avec les mêmes valeurs. Donc, si par exemple la première transformation correspond à un contraste de 20, un rotation de 10° et 50% de bruit, toutes les images subiront cette même série de traitements avec ces valeurs précises, et non de nouvelles tirées à chaque passage.

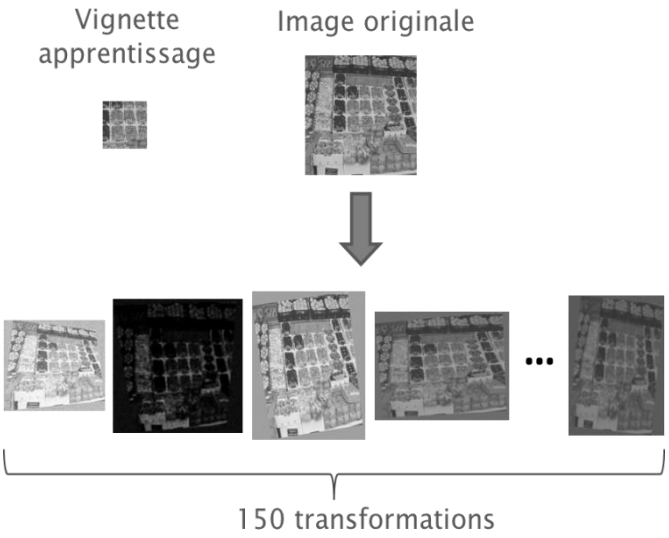


Figure III-18 Création d'exemples positifs à partir des images de la base d'apprentissage

Notre corpus d'apprentissage est donc constitué de 250 images (150 de façades, et 100 autres images variées) et notre corpus de test de 250 autres images de façades, de 350 images de films ainsi que des 250 images du corpus d'apprentissage et les 150 variations de chacune, soit au total 38350 images. Cependant chaque modèle appris ne sera testé que sur les 150 variations lui correspondant et non sur celles des 249 autres images de la base d'apprentissage. Donc chacun sera testé sur 1000 images, 151 positives, et 849 négatives, tel que résumé dans la Figure III-19.

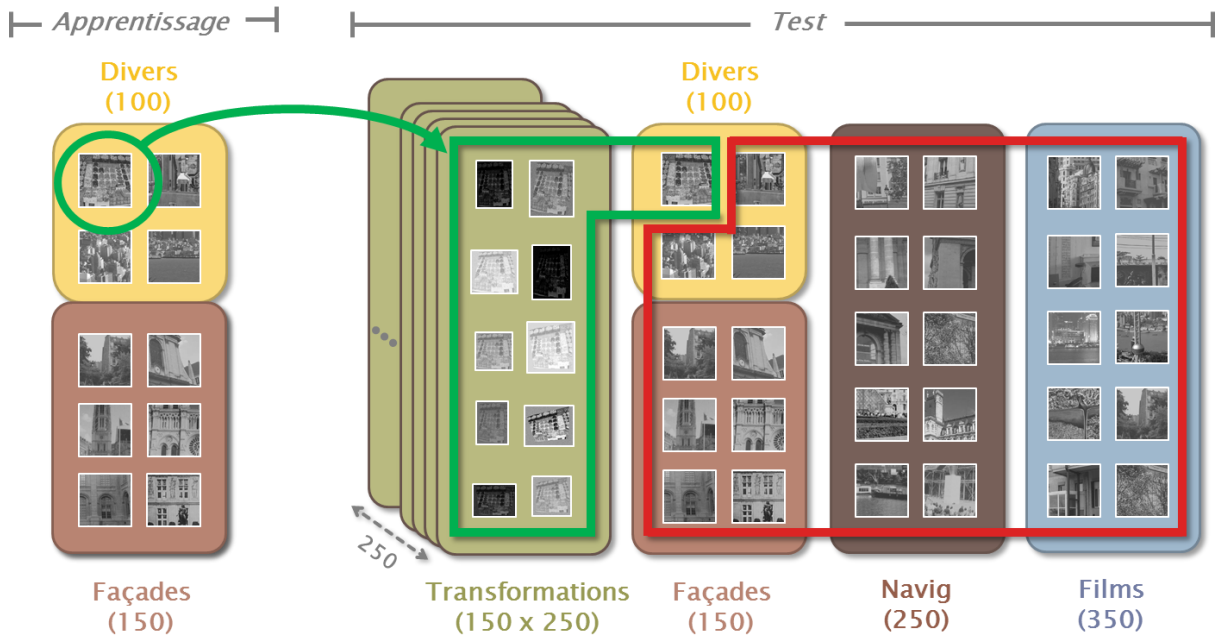


Figure III-19 Résumé des images constituant les corpus d'apprentissage et de test. Pour chaque image apprise nous avons donc 151 exemplaires positifs (en vert) et 849 négatifs (en rouge)

3.2.2 Métriques d'évaluation

Afin de clarifier la lecture des résultats présentés par la suite il convient d'introduire certains termes et métriques que nous allons utiliser. L'évaluation d'un classifieur, peu importe le domaine d'application, repose généralement sur une matrice de confusion contenant 4 valeurs, les nombres de Vrais Positifs, Vrais Négatifs, Faux Positifs et Faux Négatifs (que nous noterons par la suite *VP*, *VN*, *FP* et *FN*). Dans notre contexte d'application les exemplaires testés sont des images, dans lesquelles le moteur de reconnaissance recherche la présence de cibles apprises. Si l'algorithme effectue une détection sur une image contenant la cible il s'agira d'un *VP*, si la cible n'était pas présente, d'un *FP* (également appelé fausse alarme, ou erreur de type 1). Si en revanche aucune détection n'intervient alors qu'une cible apparaissait il s'agira d'un *FN* (ou erreur de type 2), et si elle est effectivement absente d'un *VN*. Cette classification permet d'évaluer la catégorisation de l'image comme contenant ou non le modèle testé. Cependant, l'algorithme Spikenet fournissant également les coordonnées de la cible dans l'image, nous souhaiterions évaluer ses performances de localisation. Pour cela nous allons compléter la table de confusion avec chacune des détections (ou hits) ayant eu lieu[1]. Les hits aux coordonnées

[1] Une image pouvant contenir plusieurs détections du même modèle.

correctes[1] seront comptés comme VP, les autres[2] comme FP. Chaque image contenant la cible mais n'ayant déclenché aucune détection sera enfin ajoutée aux *FN*.

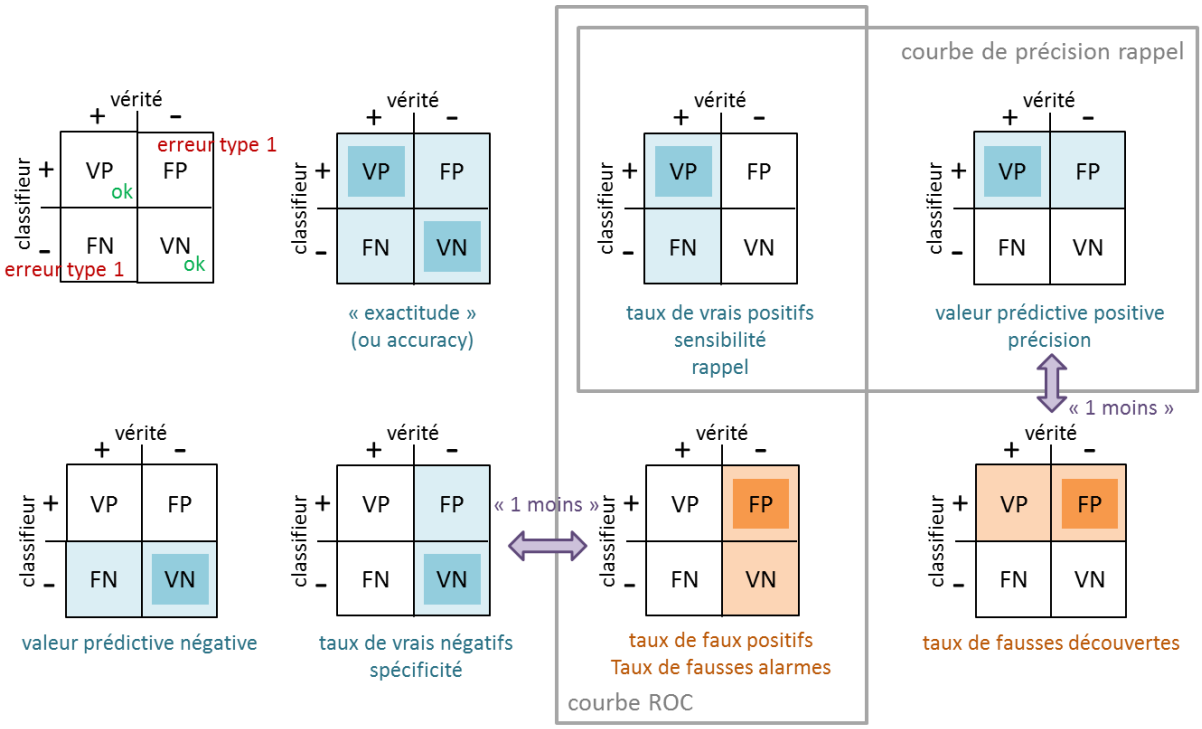


Figure III-20 Différentes mesures calculées à partir des matrices de confusion contenant le nombre de Vrais Positifs, Faux Positifs, Vrais Négatifs et Faux négatifs. Les couleurs sombres correspondent aux numérateurs, et les claires aux dénominateurs. Toutes ces métriques sont comprises entre 0 et 1, lorsque les performances du classifieur sont bonnes celles en bleu tendent vers 1 et celles en orange vers 0.

A partir de cette table de contingence nous pouvons donc dériver différentes mesures reflétant différentes propriétés du classifieur (nous noterons P le nombre d'images positives, c'est-à-dire contenant la cible considérée, soit $P = VP + FN$, et N le nombre d'images négatives tel que $N = FP + VN$). Ces différentes mesures sont résumées la Figure III-20.

- **Précision** : appelée également valeur prédictive positive, elle correspond au nombre d'images correctement classées par rapport au nombre total d'images classifiées positives, c'est-à-dire la proportion des détections étant des vrais positifs

$$Précision = \frac{VP}{VP + FP}$$

[1] Nous utilisons pour cela le critère en vigueur dans la tâche de localisation du Pascal VOC Challenge : que la distance entre la position détectée et la position réelle soit inférieure à un tiers de la taille de l'objet, de sorte que le taux de recouvrement soit d'au moins 50 %.
[2] Ceux de coordonnées incorrectes dans une image contenant la cible ainsi que l'ensemble des hits des images où elle est absente.

- **Rappel** : aussi appelé sensibilité ou taux de vrais positifs, le rappel représente le pourcentage correctement classifié parmi l'ensemble des images positives

$$Rappel = \frac{VP}{P}$$

- **Taux de faux positifs** (TFP) : c'est-à-dire le nombre de fausses alarmes parmi l'ensemble des images négatives

$$TFP = \frac{FP}{N}$$

- **Taux de fausses découvertes** (TFD), soit le nombre de fausses alarmes parmi l'ensemble des images classifiées positives

$$\mathrm{TFD} = \frac{FP}{FP + VP}$$

- **Spécificité** (notée SPC) : correspond au pourcentage d'images sans détection parmi l'ensemble des images négatives

$$SPC = \frac{VN}{N}$$

- **Valeur prédictive négative** (ou VPN) : le nombre d'images négatives parmi celles classifiées comme négatives.

$$VPN = \frac{TN}{TN + FN}$$

Certaines mesures permettent de combiner les 4 valeurs de ces tables de confusion. L'*overall accuracy,* bien que contestée reste une des plus répandues [Huang and Ling, 2005; Lavrac et al., 1999]. Elle correspond au nombre d'images correctement classifiées par rapport au nombre total d'images testées, soit $(VP + VN) / (P + N)$**.** Les autres les plus couramment utilisées sont la mesure *F1*, définie comme la moyenne harmonique du rappel et de la précision[1], et le coefficient de corrélation de Matthews, ou MCC, dont la formule est fournie ci-dessous :

$$MCC = \frac{(VP \times VN) - (FP \times FN)}{\sqrt{(VP + FP)\,(VP + FN)\,(VN + FP)\,(VN + FN)}}$$

[1] Egalement appelée F-score, ou F-mesure, il s'agit cas particulier de la mesure F^β :

$$F^\beta = (1 + \beta^2) \times \frac{precision \times rappel}{(\beta^2 \times precision) + rappel}$$

Deux autres mesures sont introduites dans [Powers, 2011], dont la moyenne géométrique[1], correspond au coefficient de Matthews : la première, nomée *Informedness* est égale à $rappel + spécificté - 1$, tandis que la seconde (*Markedness*), est définie par $précision + valeur\ prédictive\ négative - 1$. Nous pouvons pour finir, mentionner l'index de performance, ou *PI*, défini dans [Shafi et al., 2008] comme la moyenne arithmétique du rappel, de la précision, de la spécificité et de l'*overall accuracy*.

Courbes ROC et PR

Un classifieur produit généralement, pour un document particulier, un score sur l'appartenance ou non de la classe apprise. Il est donc possible d'ordonner différents documents en fonction de ces scores, puis de choisir un nombre donné d'éléments pour répondre à une requête comme, par exemple, trouver dix images de maisons parmi mille. Il est aussi possible de fixer un seuil de décision, servant de critère pour catégoriser le document comme positif ou négatif. En variant ce seuil on modifie le comportement du classifieur. Avec un seuil élevé nous obtiendrons un classifieur conservateur, le taux de faux positifs sera faible (peu d'erreurs parmi les documents déclarés positifs), mais le taux de rappel également (peu de documents positifs effectivement classés comme tels). A l'inverse en diminuant le seuil (plus libéral), ces deux valeurs augmenteront, plus de fausses alarmes, mais également plus de vrais positifs. Toutes les métriques mentionnées précédemment sont donc fonction de ce paramètre, et pour comparer des performances de classification il convient donc de considérer leurs valeurs pour différents seuils, plutôt que pour un seuil donné.

On représente donc souvent celles-ci sous forme de courbes, en prenant généralement ces mesures par paires tout en variant le seuil. Une des méthodes les plus couramment utilisé est la courbe *Receiver Operating Characteristic Curve*, ou ROC [Flach, 2003], permettant d'estimer la valeur d'un test ou d'un classifieur en traçant le rappel par rapport au taux de faux positifs [Green and Swets, 1966; Metz, 1978]. Il s'en suit que ces courbes passent inévitablement par le point $(0,0)$ correspondant au seuil maximal ou tous les éléments sont considérés négatifs, et par le point $(1,1)$ où tous sont classifiés positifs (voir Figure III-21). La diagonale entre ces deux points correspond à un classifieur sans pouvoir discriminant, au niveau de la chance, pour des tailles d'échantillons positifs et négatifs égales. Par conséquent les courbes ROC sont supposées être au-dessus de cet axe, et meilleur le classifieur sera, plus la courbe se rapprochera du point $(0,1)$, dans le coin supérieur gauche du graphe.

[1] La moyenne géométrique de n valeurs x_1 à x_n est égale à $\sqrt[n]{\sum_{i=1}^{n} x_i}$

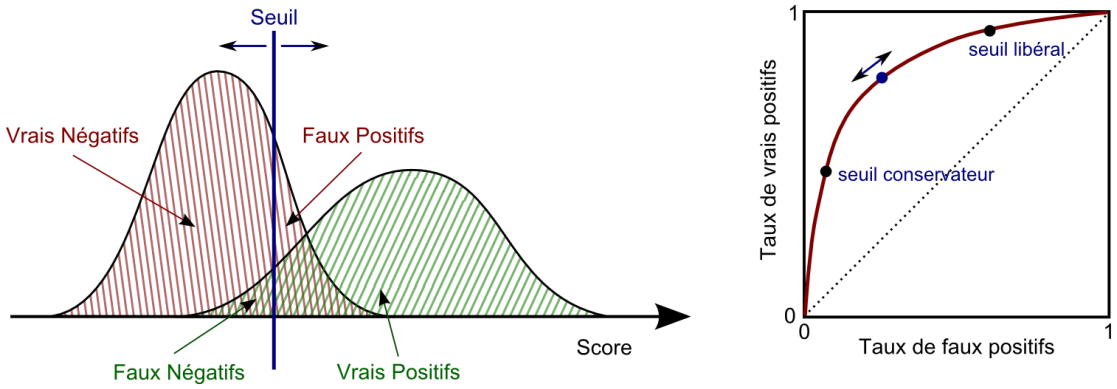


Figure III-21 A gauche, distribution des scores des éléments positifs et négatifs, et seuil de décision les catégorisant comme VP, FP, VN ou FN. A droite illustration d'une courbe ROC standard.

Pour comparer deux classifieurs, il est possible de fixer une valeur souhaitée comme taux de faux positifs, puis d'observer pour chacun le taux de vrais positifs correspondant (l'inverse est évidement aussi possible, en fixant le rappel). Mais si dans certains cas un classifieur dominera un autre en tous les points de la courbe, dans d'autres celui présentant les meilleurs résultats pourra varier selon la valeur considérée (les courbes se croisant). Une méthode plus générique, reflétant leur comportement global sur l'ensemble des seuils possibles, consiste donc à comparer l'aire sous la courbe des deux classifieurs [Bradley, 1997; DeLong et al., 1988]. Il a été montré que dans la plupart des cas d'application des algorithmes d'apprentissage, cette mesure s'avérait être la plus informative et la plus pertinente pour l'optimisation et la sélection de modèles, notamment en comparaison à l'*accuracy* [Fatourechi et al., 2008; Powers, 2012; Provost and Fawcett, 1997].

Les courbes ROC ont été très largement étudiées dans le champ des diagnostics médicaux depuis les années 70, et de nombreuses méthodes paramétriques ont été développées afin d'estimer l'aire sous la courbe[1] pour un faible nombre de valeurs possibles de jugement. Parmi celles-ci la technique la plus couramment employée est l'estimation de l'aire par maximum de vraisemblance, sous l'hypothèse d'une distribution binormale [Dorfman and Alf, 1969; Metz et al., 1984]. Cependant dans le domaine de la vision artificielle, si les scores de classification sont aussi souvent discrétisés, il est possible de générer un nombre de points beaucoup plus important pour le tracé des courbes ROC, et la méthode trapézoïdale s'avère donc une très bonne approximation [Faraggi and Reiser, 2002], même si elle sous-estime nécessairement légèrement l'aire réelle [Hanley and McNeil, 1982]. Cette méthode non-paramétrique de calcul des aires sous les courbes ROC a été démontrée équivalente au test de permutations [Sheskin, 2000], au test signé des rangs de

[1] Généralement notée AUC pour *Area Under Curve*

Wilcoxon [Mason and Graham, 2002], ainsi qu'au test de Mann-Whitney[1] [Bamber, 1975]. En résumé l'aire sous une courbe ROC peut être interprétée comme la probabilité d'une décision correcte dans un choix forcé à deux alternatives [Green and Swets, 1966]. Pour comparer deux AUC il est nécessaire de calculer la significativité de leur différence, et donc pour cela leur variance. Différentes méthodes ont été proposées dans ce but [Hanley and Hajian-Tilaki, 1997], les plus répandues reposent sur les propriétés de la statistique de Wilcoxon [Hanley and McNeil, 1982], sur la statistique U de Mann-Whitney [DeLong et al., 1988], ou sur des techniques de *jackknifing*[2] [McNeil and Hanley, 1983].

Notons que les courbes ROC, parce qu'elles illustrent les taux de faux positifs et de vrais positifs, sont insensibles au pourcentage d'éléments positifs et négatifs dans l'échantillon [Metz, 1978; Powers, 2011]. Pour des distributions variables, il est donc souvent suggéré d'utiliser les courbes de Précision-Rappel (PR) [Fawcett, 2004]. Celles-ci sont similaires au courbes ROC mais permettent de visualiser, comme leur nom l'indique, la précision (en ordonnée) en fonction du rappel (en abscisse). La Figure III-22, tirée de [Fawcett, 2006], illustre ce phénomène, les courbes ROC et PR de deux classifieurs ont été calculées pour deux échantillons, le premier ayant une distribution de classes balancée (autant d'éléments positifs que négatifs), le deuxième comprenant dix fois plus d'exemples négatifs. Les courbes ROC, insensibles à ces distributions, restent donc inchangées, alors qu'on observe de larges différences dans les courbes PR.

Il existe une équivalence directe entre chaque point de l'espace ROC et de l'espace PR. En revanche l'interpolation entre ces points est différente : pour les courbes ROC une interpolation linéaire suffit alors que les courbes PR nécessitent une méthode non-linéaire plus complexe, comme l'*Achievable PR Curve*, proposée dans [Davis and Goadrich, 2006] permettant la création de points intermédiaires nécessaires au calcul de l'aire sous courbe (AUC-PR). Une métrique plus simple, appelée *Average Precision,* est donc souvent utilisée, comme dans le Pascal Visual Object Classes Challenge [Everingham et al., 2009] depuis 2007. Elle consiste à moyenner une précision interpolée pour 11 valeurs de rappels uniformément espacées entre 0 et 1. Cette interpolation consiste à prendre la valeur maximale de précision pour un rappel supérieur ou égal à celui considéré, la courbe étant (généralement) décroissante.

[1] Le test de Mann-Whitney calcule la probabilité qu'un élément tiré au hasard parmi un échantillon A soit supérieur à celui d'un échantillon B, également tiré au hasard [Mann and Whitney, 1947].
[2] Le *jackknifing* est une méthode de ré-échantillonnage statistique proche du *bootstrap.*

L'utilisation préférable d'une courbe plutôt qu'une autre n'est pas une question tranchée. D'autant que souvent les résultats produits sont similaires, car comme démontré dans [Everingham et al., 2009] une courbe domine une autre dans l'espace ROC si et seulement si elle la domine également dans l'espace PR. Dans notre cas nous avons généré et étudié les deux types, et au vu des comportements semblables avons retenu les courbes ROC pour laquelle la méthode de calcul de l'AUC est plus directe.

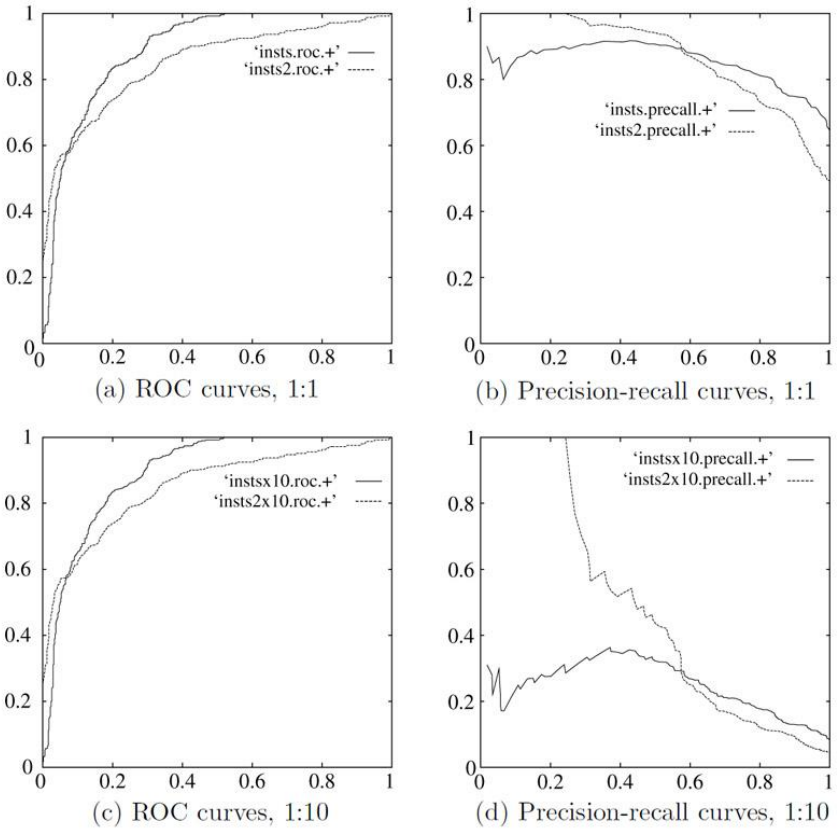
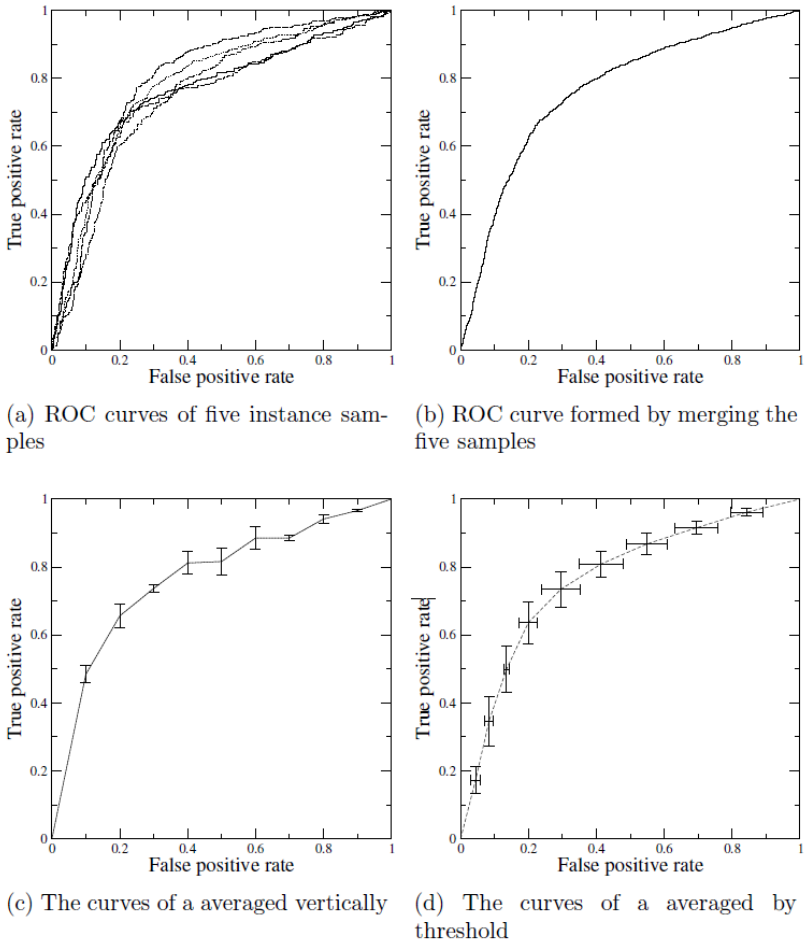


Figure III-22 Courbes ROC et PR de deux classifieurs sur deux échantillons : l'un ayant autant de positifs que de distracteurs (1:1), le second avec 10 fois plus de distracteurs (1:10). Figure tirée de [Fawcett, 2006].

Combinaison de courbes ROC

Notons enfin que comme décrit précédemment, ces mesures permettent de caractériser un classifieur binaire. Dans le cas de Spikenet de nombreux modèles (motifs visuels préalablement appris) sont conjointement testés sur les nouvelles images fournies en entrée. Nous considérerons chacun de ces modèles comme un classifieur binaire pour calculer différents scores propres à chacun, qui pourront ensuite être moyennés afin d'obtenir des mesures globales sur l'ensemble des motifs appris.

Pour tracer une courbes ROC globale représentant un ensemble de classifieurs, il s'agit donc de combiner leur réponse. Il existe pour cela trois grande méthodes, décrites notamment dans [Bradley, 1997; Fawcett, 2004; Macskassy and Provost, 2004], et illustrées dans la Figure III-23.



(a) ROC curves of five instance samples

(b) ROC curve formed by merging the five samples

(c) The curves of a averaged vertically

(d) The curves of a averaged by threshold

Figure III-23 Méthodes de moyennage des courbes ROC : 5 courbes, présentées dans (a), sont combinées par *pooling* (b), moyennage vertical (c) ou moyennage par seuil (c) ; tirée de [Bradley, 1997; Fawcett, 2004]

La première, nommée *pooling*, consiste à regrouper l'ensemble des réponses de tous les classifieurs (labels et score de classification), puis de générer une courbe ROC traditionnelle de ce méta-classifieur. Si elle est relativement simple à mettre en place, cette technique ne permet cependant pas d'estimer la variance des mesures, et fait de plus l'hypothèse de seuils d'utilisation similaires pour chacun des classifieurs, ce qui n'est pas nécessairement le cas. Les deux autres méthodes, consistant à moyenner les résultats, permettent de dépasser ces limitations. La première, par moyennage vertical, consiste à extraire et moyenner les taux de vrais positifs (en ordonnée) pour des valeurs données de faux positifs (en abscisse). Néanmoins, la variable indépendante à partir de laquelle le

moyennage est réalisé (le taux de faux positifs) n'étant pas directement contrôlable, il est généralement recommandé de moyenner en fonction des seuils (méthode que nous avons adoptée). Les échantillons ne reposent donc pas sur la position des points dans l'espace ROC, mais sur les seuils qui produisent ceux-ci. En pratique il suffit donc de choisir un ensemble de seuils uniformément distribués, puis de calculer pour chacun les valeurs de vrais et faux positifs des différents classifieurs, dont les moyennes fourniront les points de la courbe finale. Pour ce qui est des aires sous courbes (AUC), l'approche généralement employée, que nous utiliserons par la suite, consiste à calculer les aires pour l'ensemble des classifieurs pris individuellement, puis d'extraire la moyenne de celles-ci, ainsi que les intervalles de confiance à 95%, grâce à la méthode de bootstrap non paramétrique BCA (*Bias Corrected and Accelerated percentile method*) présentée dans [DiCiccio and Efron, 1996].

3.2.3 Plateforme expérimentale

L'algorithme Spikenet MutltiRes a été implémenté à partir du noyau Spikenet originel, au sein d'une librairie C++ dont les différentes primitives pour l'apprentissage et la détection de cibles sont restées semblables au noyau originel de sorte que les 2 versions soient interchangeables. Seules certaines primitives supplémentaires ont été ajoutées dans le but de pouvoir modifier les différents paramètres spécifiques à cette nouvelle architecture tels que le nombre et les dimensions des différentes résolutions, le nombre de poids pour chacune, les stratégies de vote entre échelles,…

Un programme avec une interface graphique a également été développé afin d'exploiter ce noyau. Celui-ci, également implémenté en C++ et utilisant la librairie QT4 permet l'automatisation des évaluations en lançant l'apprentissage sur une liste d'images, et la détection sur une autre. Il permet également de visualiser les modèles appris, de fixer les différents paramètres de l'apprentissage et de la détection, ainsi que de sauvegarder tous les résultats dans des fichiers CVS. Ceux-ci contiennent l'ensemble des détections et leurs différents attributs (modèle déclenchant le hit, score de celui-ci, image testée, coordonnées du hit, temps de traitement, etc.) qui ont ensuite été exploités sous Matlab pour l'analyse et la visualisation de ces résultats. La création des corpus d'apprentissage et de tests consistant à redimensionner et normaliser l'ensemble des images, ainsi que de générer les 150 transformations mentionnées précédemment, ont également été effectuées au moyen d'un ensemble de scripts Matlab. Se reporter aux annexes pour un aperçu des programmes réalisés.

3.3 Résultats

Afin de mettre en place des stratégies combinant l'utilisation de différentes résolutions, il est naturellement nécessaire d'identifier leurs spécificités. Nous nous sommes donc dans un premier temps intéressés aux performances et propriétés de chacune des résolutions prises indépendamment.

3.3.1 Nombre de poids par échelle

Comme nous l'avons observé dans l'étude préliminaire (section III.3.1.2), en adoptant durant l'apprentissage des mécanismes de compétition inter-échelles, la répartition des neurones dans chacune des résolutions est très proche d'une fonction linéaire de leur taille. Lorsque le nombre global de poids est trop faible, nous avons constaté que la redondance entre les différentes échelles entraînait une perte d'information sur le stimulus appris, mais qu'elle disparaissait cependant au-delà d'un certain seuil. Une augmentation relative de ce nombre total de poids (de sorte que chaque résolution puisse conserver au moins 1/3 de ses réponses, tel que fixé dans l'algorithme mono-résolution) n'ayant que très peu d'impact sur les temps de traitement, nous avons fait le choix d'utiliser une méthode d'apprentissage appliquée indépendamment à chaque échelle, pour un nombre fixe de spikes (proportionnel à la résolution), permettant ainsi de simplifier l'architecture et l'implémentation.

L'analyse des performances de classification pour différents nombres de poids à chacune des résolutions[1] nous a permis de conclure qu'au-delà d'un certain nombre, relativement bas[2], ceux-ci n'avaient presqu'aucune incidence sur les résultats, provoquant simplement une translation des niveaux d'activation et donc un décalage des seuils à appliquer pour conserver des performances similaires (cf. Figure III-24). Au vu de ces données, nous avons décidé de conserver le même ratio de neurones retenus lors de l'apprentissage que celui utilisé dans le noyau Spikenet originel (soit environ 1/3 des poids), que nous avons appliqué à chacune des résolutions en fonction de leur taille (à l'exception des 3 plus basses, pour lequel il a été fixé manuellement).

[1] Dont les résultats pour trois d'entre elles sont présentés dans la Figure III-24.
[2] 1 Une dizaine pour la résolution 8 x 8, 25 un trentaine pour 12 x 12, puis seulement de 50 à 100 au-delà.

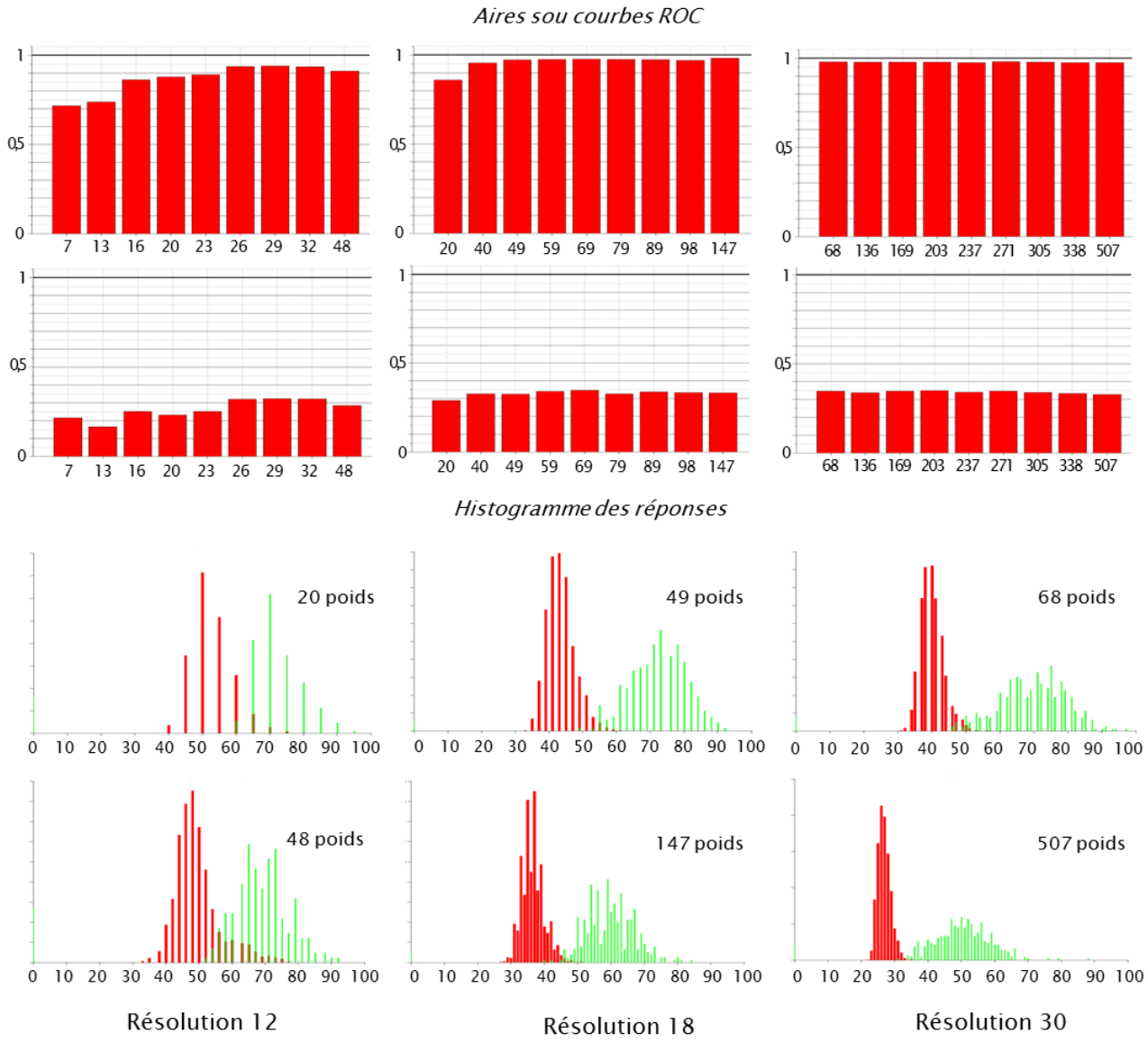


Figure III-24 Effet du nombre de poids retenus par échelle. Trois échelles (12, 18 et 30) sont présentées, une par colonne. Pour chacune, l'aire sous courbe ROC est fournie en fonction du nombre de poids retenus, sur deux échantillons de tests (l'un simple, l'autre difficile). Des histogrammes présentant l'activation moyenne des cibles et distracteurs sont également proposés pour deux nombres de poids différents afin d'illustrer le décalage des seuils d'activation.

3.3.2 Performances de classification

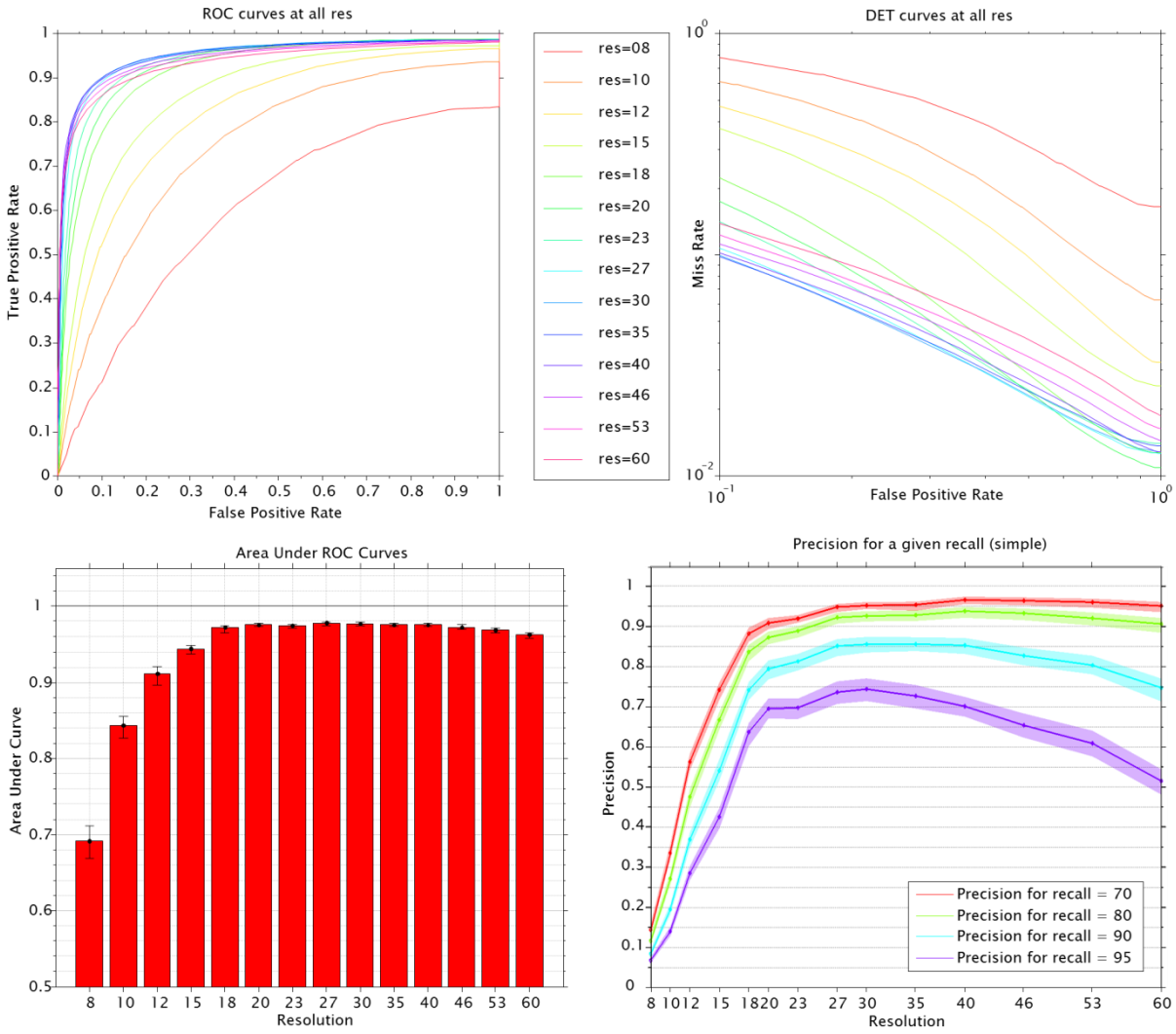


Figure III-25 Performance de classification par résolution. Courbes ROC et DET sur la première ligne. Aire sous courbe ROC et précision moyenne pour des valeurs fixes de rappel sur la deuxième (avec intervalles de confiances à 95 %)

Les courbes précisions rappels et l'aire sous celles-ci (AUC) fournissent, comme nous l'avons mis dans la section III.3.2 (Méthodes), une bonne estimation de la performance d'un classifieur sur l'ensemble de seuils d'utilisation. Celles-ci sont présentées pour chacune des résolutions dans la Figure III-25, conjointement à des courbes DET[1], permettant une comparaison visuelle plus fine des différences aux seuils critiques [A. Martin et al., 1997],

[1] Les courbes DET (*Detection Error Tradeoff*) illustrent le taux de « manqués » (*MissRate*), correspondant au nombre de faux négatifs sur l'ensemble des éléments positifs ($MissRate = FN/(VP + FN)$ c'est-à-dire $1 - Rappel$), en fonction du taux de faux positifs, sur des échelles logarithmique. Contrairement aux courbes ROC et PR, les meilleures performances dans les courbes DET correspondent aux courbes les plus basses (présentant le moins d'erreur).

ainsi qu'à la précision moyenne des classifieurs pour différentes valeurs de rappel. Ces données montrent des résultats très faibles aux résolutions les plus basses, pour lesquelles les modèles ne possèdent pas suffisant d'information sur le motif visuel, qui s'améliorent ensuite rapidement jusqu'à 18 px pour finalement se stabiliser, avec un léger pic autour des échelles 27/30. Il est important de noter que si l'on continue d'augmenter la taille d'apprentissage, les performances, elles, ne vont pas nécessairement s'accroitre. On observe d'ailleurs même une baisse de classification entre 27 et 60 px, particulièrement visible sur les valeurs de précision. Ce phénomène s'explique par le fait que si, certes, on possède plus de détails sur l'objet à apprendre, des informations trop fines vont avoir un impact sur la tolérance aux transformations. Cette sensibilité aux changements locaux se traduit donc par une baisse de la tolérance du classifieur.

Ajustement de la difficulté du jeu de données

Les résultats que nous venons d'analyser reflètent certaines tendances, cependant parmi les différentes résolutions, nombreuses sont celles dont les performances atteignent un maximum, correspondant à une courbe ROC se rapprochant du coin supérieur gauche, pour laquelle l'AUC tend donc vers 1. Un effet « plafond » pourrait par conséquent gommer certaines différences de comportement entre des résolutions aux AUC les plus élevées. Nous avons donc décidé de constituer plusieurs jeux de données de difficulté variable. Comme nous l'avons détaillé dans la section 3.2.1, le corpus de test initial a été construit en appliquant des combinaisons de paramètres aléatoires pour une dizaine de transformations géométriques et photométriques. Les scores de classification sur cette première base étaient extrêmement élevés, supérieurs même à ceux présentés plus tôt, avec des AUC au-dessus de 0.98 pour toutes les résolutions à partir de l'échelle 12. Nous avons par conséquent décidé d'augmenter les plages de valeurs dans lesquelles étaient tirés les paramètres aléatoires pour chacune des transformations. La taille était par exemple contrainte entre 90 et 110 %, nous sommes passé à 75-125 %, la quantité de bruit de 0 à 50 %, a été élargie à 0-100 %, etc. Cependant en utilisant des valeurs trop extrêmes, l'image se trouve tellement modifiée que les modèles ne permettront jamais de détection, peu importe leur tolérance, ce qui a conduit à des taux de rappel très bas et donc des performances générales trop faibles pour juger de la qualité des différents classifieurs.

En observant les résultats de l'évaluation de l'effet des différentes transformations prises individuellement sur un spectre très large de paramètres, nous avons pu ajuster une fonction prédisant pour chacune la quantité de perte de signal moyenne selon la valeur choisie. Combinées dans un modèle paramétrique, nous avons alors comparé ces prédictions avec les réponses des différents classifieurs sur les corpus construits précédemment. Ces résultats, présentés dans la Figure III-26, montrent effectivement une corrélation importante, permettant de valider ce modèle.

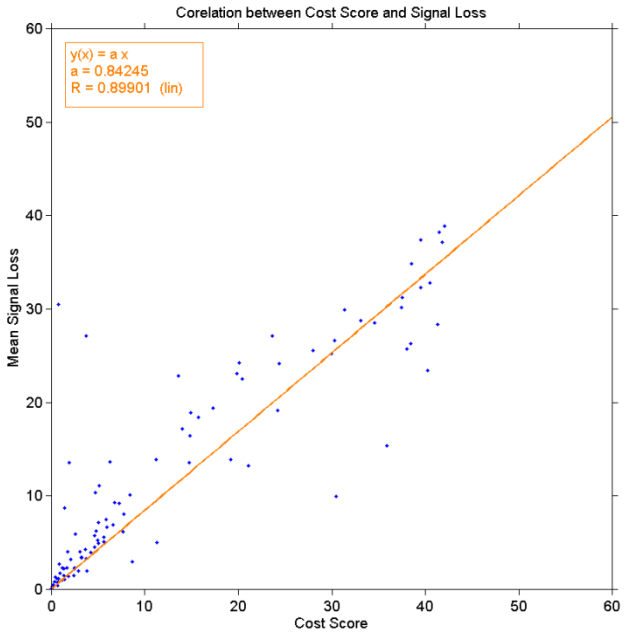


Figure III-26 Perte moyenne de signal de l'ensemble des modèles (toutes résolutions confondues), en fonction de la prédiction du cout des transformations appliquées.

En utilisant cette fonction de prédiction de l'impact des combinaisons de transformation, nous avons alors pu construire deux nouveaux sets, de difficultés variables (présentés dans la Figure III-27). Pour cela, au lieu de tirer aléatoirement les paramètres de chaque transformation, nous avons défini un intervalle de perte de signal souhaitée (entre 0 et 50 par exemple), permettant d'accroître la difficulté en augmentant ses valeurs. Finalement pour définir les valeurs exactes de chaque combinaison de transformations, une valeur aléatoire est tirée dans l'intervalle choisi, puis une répartition de cette diminution de la réponse parmi les différentes transformations est elle aussi tirée aléatoirement[1]. Pour finir, en multipliant leurs contributions avec la perte totale attendue, nous obtenons la baisse de réponse désirée pour chacune, permettant de déterminer grâce au modèle de prédiction le paramètre de transformation aboutissant à la valeur s'en rapprochant le plus.

[1] On tire pour cela autant de valeurs que de transformations, normées en les divisant par leur somme.

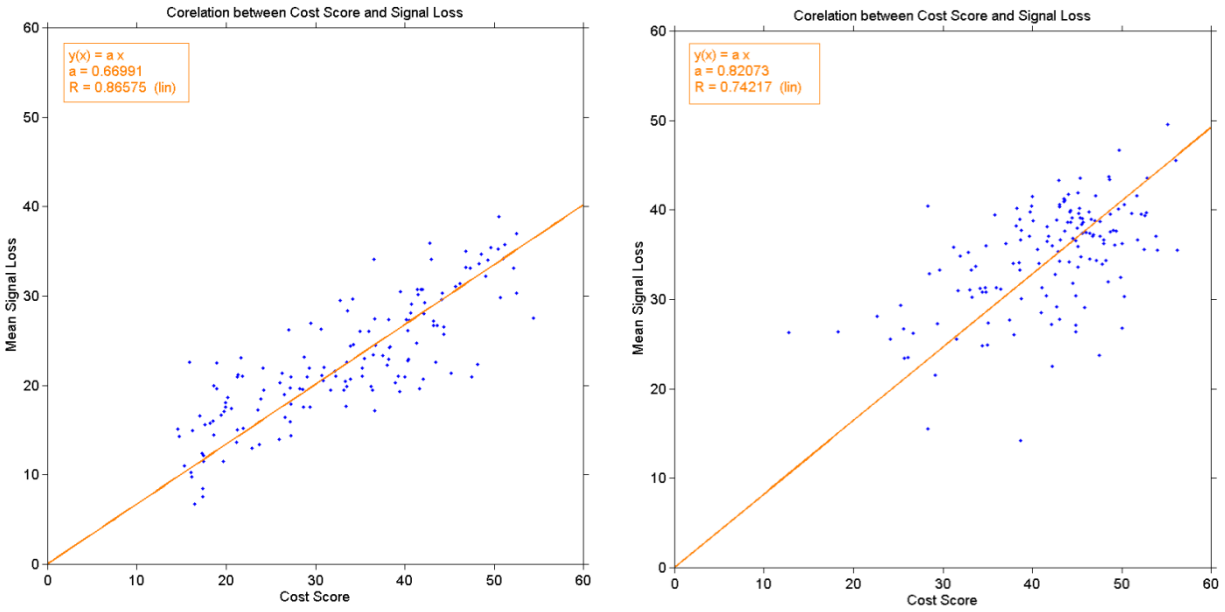


Figure III-27 Perte de signal observée sur les deux corpus générés en utilisant le modèle de prédiction du coût des transformations appliquées. Le graphique de gauche correspond à la base de test de difficulté moyenne, celui de droite à la base difficile.

Les résultats obtenus sur le premier corpus généré par ce procédé, de difficulté moyenne, ont été présentés précédemment. Ceux du second, de difficulté accrue, sont fournis dans la Figure III-28. Comme nous le supposions, des scores plus bas permettent de dégager de nouvelles tendances masquées par l'effet plafond, comme la baisse plus conséquente des performances aux hautes résolutions. Dans les résolutions intermédiaires on observe également certaines différences avec l'évaluation des performances sur la première base de test, dues aux plus fortes variations des exemplaires positifs résultant de transformations affines plus importantes. On peut en effet constater que plus le taux de rappel augmente, plus les résolutions présentant la précision maximale[1] diminuent, confirmant l'intuition que des échelles plus faibles permettent un traitement plus tolérant aux modifications dans l'apparence des motifs recherchés. Cet effet se traduit non seulement sur la valeur prédictive positive, mais aussi plus globalement sur l'aire des courbes ROC, donc les maximums se situent autour des résolutions 18 et 20 (contre 27 et 30 précédemment).

[1] C'est à dire le pourcentage de vrais positifs parmi les hits du classifieur.

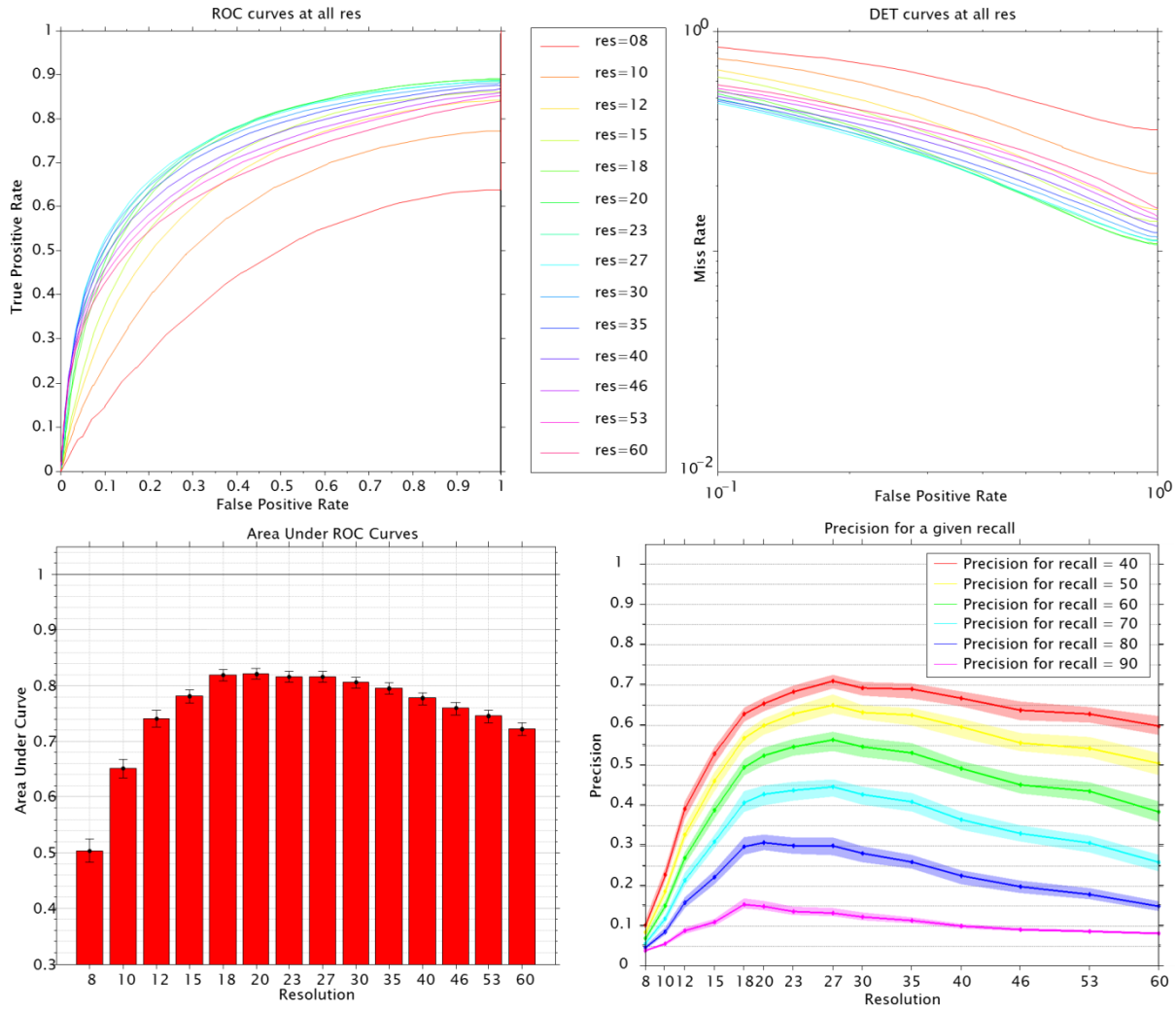


Figure III-28 Performance de classification par résolution sur la base de test de difficulté supérieure. Courbes ROC et DET sur la première ligne. Aire sous courbe ROC et précision moyenne pour des valeurs fixes de rappel sur la deuxième (avec intervalles de confiances à 95 %)

3.3.3 Temps de traitement

Chaque modèle ayant sa propre carte de propagation (basée sur les votes des poids du modèle dans l'image filtrée), il s'ensuit que le temps de traitement, pour une image donnée, est directement proportionnel au nombre de modèles testés[1]. Il est fait exception du temps de filtrage initial, indépendant de la quantité de modèles, mais qui reste négligeable en comparaison au temps de propagation. Ce phénomène apparaît dans la Figure III-29, où l'on voit que pour chaque résolution, la courbe représentant le temps de traitement en fonction du nombre de modèle est linéaire.

[1] Pour des modèles de même taille et résolution.

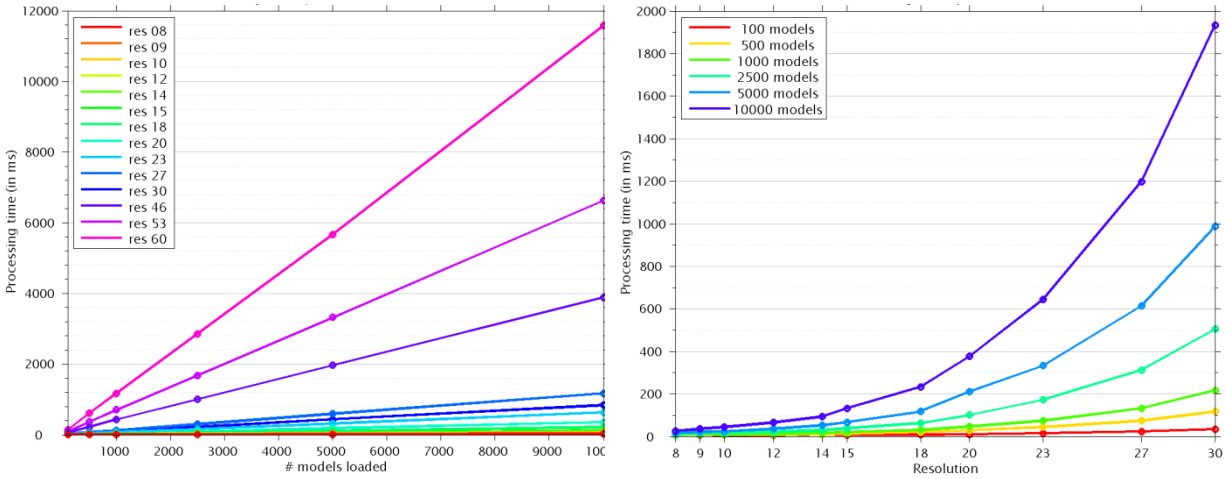


Figure III-29 Temps de traitement en fonction du nombre de modèles testés et de la résolution

Les différences entre résolutions sont quant à elles fonction du carré du nombre de pixels des patchs d'apprentissage (voir Figure III-30). Si pour un faible nombre de modèles elles restent relativement faibles de façon absolue (n'excédant par une centaine de millisecondes jusqu'à environ 50 modèles testés, quelle que soit la résolution entre 8 x 8 et 60 x 60), elles s'avèrent en revanche déterminantes lorsque le nombre de modèles augmente. A titre d'exemple, il est plus de 20 fois plus rapide d'effectuer des détections à la résolution 14 qu'à 30, celle originalement utilisée par Spikenet. Pour 10.000 modèles, il ne faudra donc qu'environ 100 ms à 14 x 14, mais près de 2 secondes à 30 x 30, un gain qui illustre l'intérêt d'utiliser de plus faibles résolutions pour limiter les temps de traitement. Sachant que tester un modèle à différentes échelles et orientations équivaut dans l'architecture Spikenet à multiplier les instances de celui-ci pour couvrir l'ensemble des combinaisons possibles de transformations, le gain de temps sera donc conséquent, même pour un nombre moins important de modèles de référence.

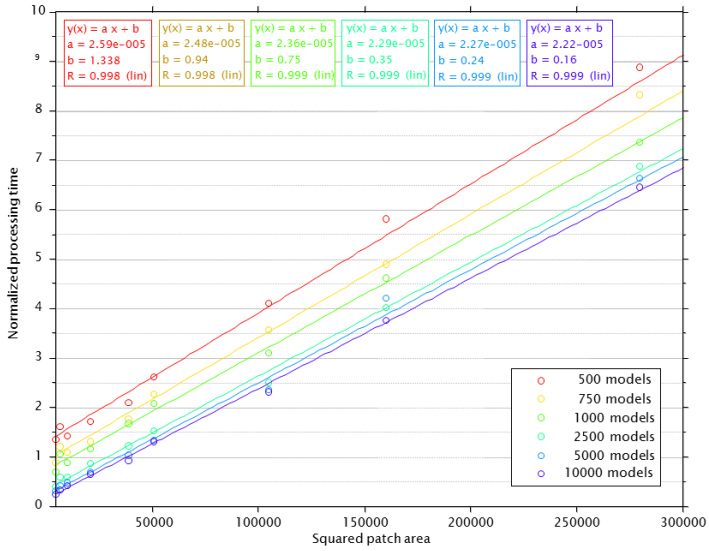


Figure III-30 Relation linéaire entre le cube de la résolution (soit le carré de l'aire des vignettes) et les temps de traitements (ceux-ci ont été normalisés en les divisant par le nombre de modèles testés)

3.3.4 Tolérance aux transformations

L'algorithme Spikenet, procède à un filtrage de l'information visuelle par des neurones sélectifs à des arêtes de différentes orientations. De plus, seules les réponses les plus rapides (correspondant donc aux saillances les plus importantes) sont conservées, et ce dans un schéma binaire (neurones actifs ou silencieux) plutôt qu'en retenant leur valeur d'activation exacte. Ces mécanismes permettent une invariance presque totale à la plupart des transformations photométriques (changements de luminance, de contraste, de gamma[1]), contrairement aux algorithmes d'apprentissage reposant sur l'intensité des pixels. On retrouve donc logiquement cette robustesse à chacune des échelles, les réponses moyennes n'étant presque pas modifiées, quels que soient les paramètres de transformation (voir Figure III-31). Le codage en orientations dominantes permet aussi une très bonne robustesse à l'ajout de bruit et de flou, pour l'ensemble des résolutions (bien que les plus hautes se montrent légèrement plus sensibles). Il supporte ainsi jusqu'à 80% de bruit ou des flous gaussiens allant jusqu'à un vingtième de la taille de l'image[2].

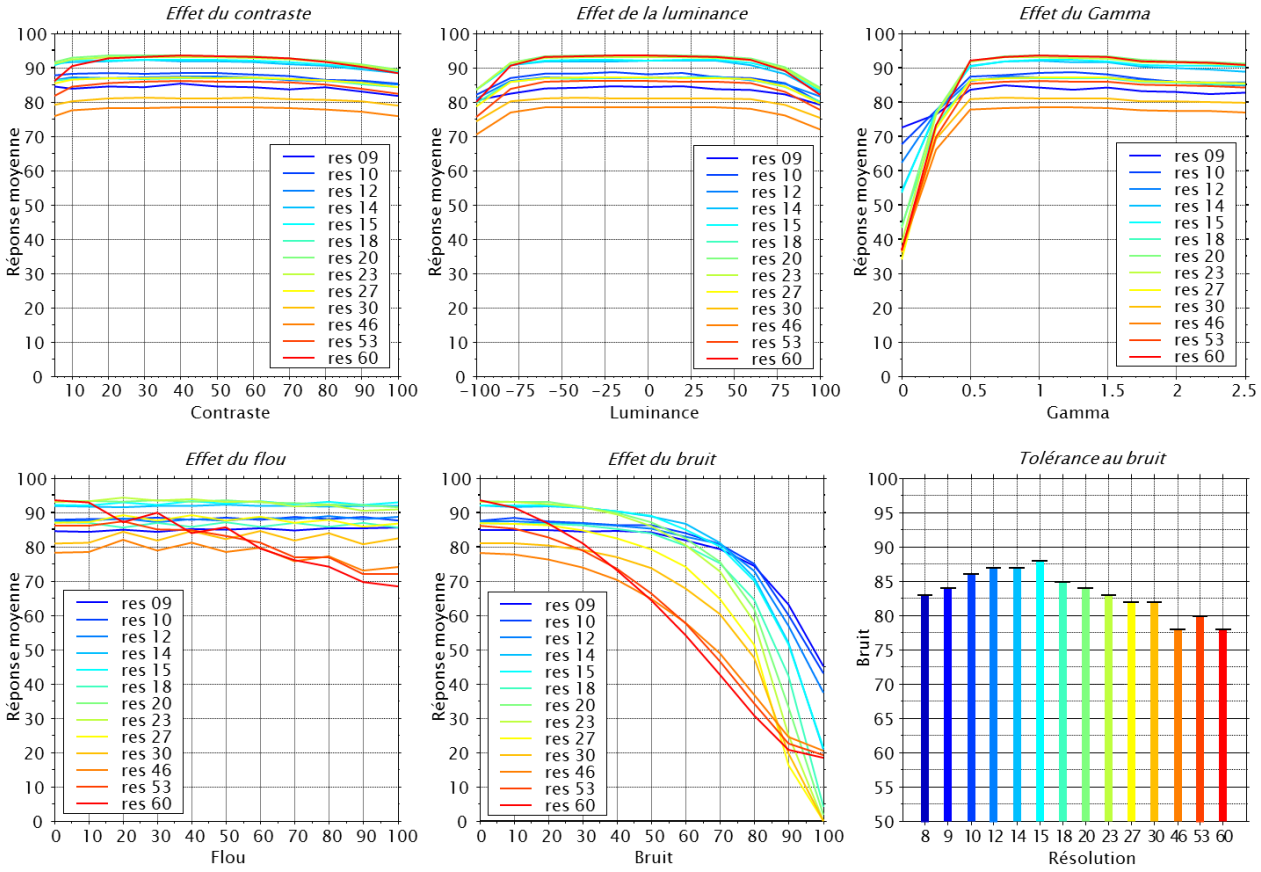


Figure III-31 Tolérance aux transformations photométriques (pour chacune des résolutions)

[1] A l'exception de valeurs extrêmement basses du gamma, supprimant presque toute l'information visuelle de l'image (qui se retrouve presque binarisée pour des valeurs proches de 0).
[2] Correspondant sur notre base de test à des fenêtres de 30 par 30 px.

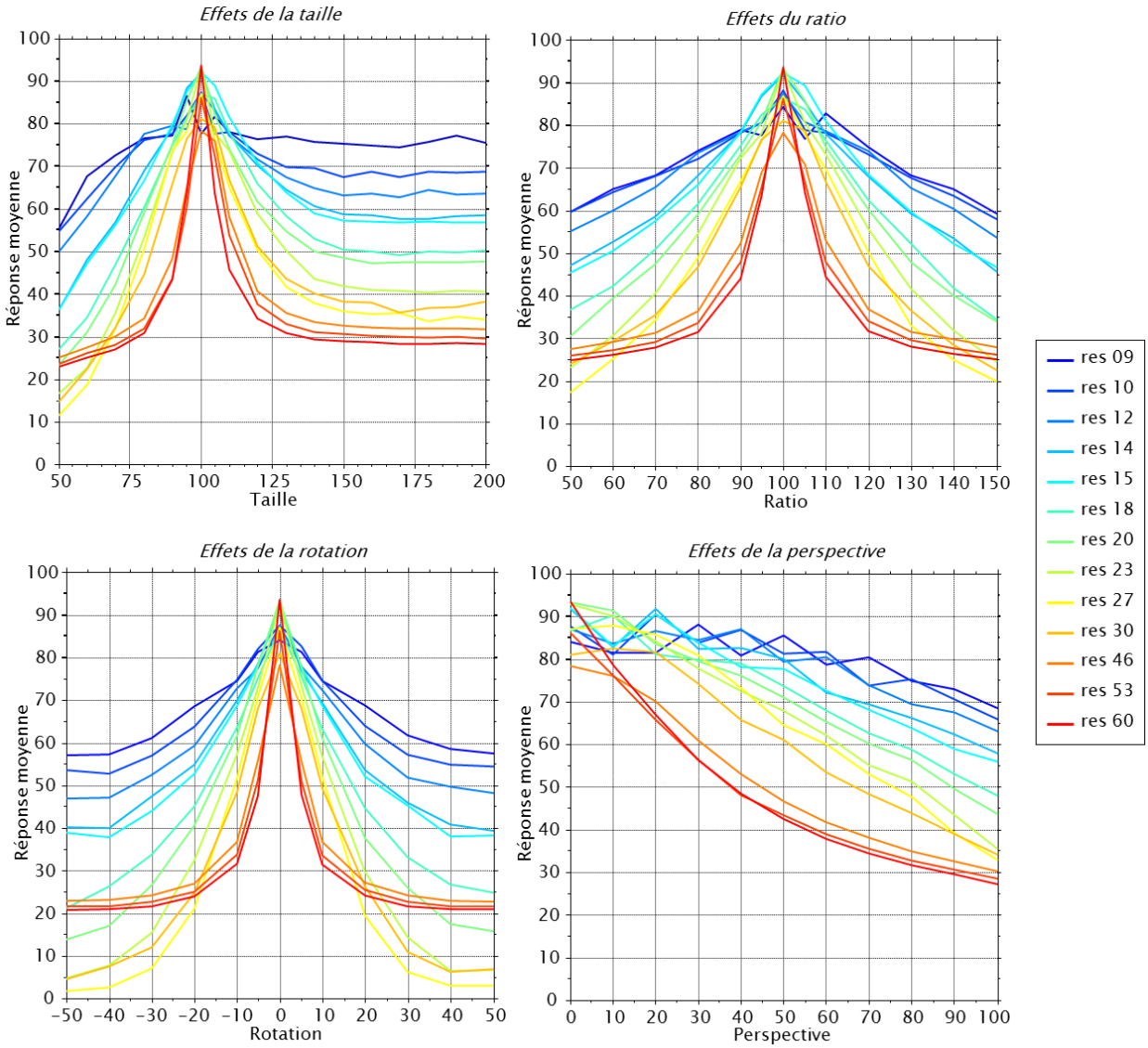


Figure III-32 Effets des transformations affines sur la réponse moyenne des modèles à chaque résolution

Les transformations affines ont en revanche un impact bien net, tel qu'en témoignent les moyennes des réponses à chacune des échelles (Figure III-32). La reconnaissance d'une forme, se basant sur la configuration spatiale d'arêtes au sein d'une fenêtre de taille fixe, est par conséquent très sujette aux changements de cette organisation qui interviennent lorsqu'on modifie le ratio de l'image, sa taille, son orientation, ou encore en appliquant des déformations perspectives. Notons au passage que nous n'avons pas testé la translation car l'architecture de l'algorithme Spikenet est par nature insensible à la position de l'objet dans l'image[1]. Elle l'est en revanche pour la rotation et le facteur d'échelle, bien qu'elle intègre des mécanismes complémentaires permettant de contourner l'absence de codage explicite

[1] Ce qui n'est pas, rappelons-le, le cas de tous les systèmes de vision artificielle.

de l'invariance à ces transformations comme c'est le cas pour les descripteurs SIFT par exemple [Lowe, 2004]. Au sein de la librairie Spikenet, chaque modèle appris comporte en effet des paramètres correspondant à la taille minimale et maximale. Par défaut tous deux égaux à 100 (le modèle n'est alors testé que pour sa taille native), ils peuvent être modifiés à la demande, ce qui conduira dans la phase de détection à appliquer la recherche de ce modèle à des versions sur et/ou sous-échantillonnées de l'image testée afin de couvrir la place d'échelle désirée. Un procédé similaire existe pour les rotations, générant différents modèles de la même cible par pas ajustables (de 5° par exemple) à l'intérieur d'un intervalle donné. Si ces méthodes permettent de couvrir virtuellement toutes les valeurs possibles de taille et d'orientation, elles nécessitent en pratique la recherche de « pseudo-modèles » générés dynamiquement, dont le nombre augmente avec la tolérance souhaitée, multipliant par conséquent les temps de traitement. Nous nous sommes donc intéressés ici seulement à la sensibilité des modèles unitaires, sans prendre en compte ces mécanismes additionnels. Ils seront évidemment intégrés à l'architecture finale, mais plus les gains en robustesses des traitements de base seront importants, plus le nombre de « pseudo-modèle » diminuera (si la tolérance aux rotations passe de 10 à 20°, on divisera ainsi par deux le nombre de modèles nécessaires pour couvrir l'ensemble des rotations possibles).

Comme le montre la Figure III-32, lorsqu'on augmente la magnitude des transformations affines appliquées à l'image, la réponse des modèles diminue d'autant plus fortement que la résolution augmente. Cette tolérance accrue des faibles résolutions doit néanmoins être mise en relation avec la distribution des réponses (aussi bien des vrais positifs que des fausses alarmes). En effet, les seuils optimaux (séparant au mieux les cibles et distracteurs) dépendent fortement du nombre de poids totaux utilisés, et donc de l'échelle. Pour estimer les plages de valeurs de tolérance à chaque résolution il est donc nécessaire de prendre en compte des seuils d'utilisation propres à chacune. Nous avons pour cela proposé deux méthodes. La première considère la moyenne des fausses alarmes plus 1,96 écart-type. Cette valeur se base sur les propriétés de la loi normale. D'après cette loi de densité, la probabilité qu'un élément appartienne à l'intervalle $[-1{,}96\,\sigma\,;1{,}96\,\sigma]$ centré sur la moyenne est de 95 %. Les distributions des distracteurs et des cibles étant toutes deux gaussiennes, l'utilisation de ce seuil revient à ne retenir que les hits ayant moins de 5 % de chances d'appartenir aux fausses alarmes. Comme on peut le voir sur la courbe de la Figure III-33, cette valeur (comme la moyenne des réponses des distracteurs), décroît régulièrement avec la résolution. Pour les échelles les plus faibles, elle s'avère néanmoins trop haute. Les basses échelles présentent en effet des sensibilités beaucoup plus faibles (les distributions des cibles et des distracteurs sont très proches), mais en même temps des valeurs de rappel plus hautes. Il est donc plus adapté de considérer le seuil optimal de fonctionnement dépendant des caractéristiques de la résolution considérée. Nous avons pour cela choisi les valeurs maximisant la mesure F1 (égale pour rappel à la moyenne harmonique de la

précision et du rappel), estimées sur les deux corpus de test présentés dans la section III.3.3.2. Les tolérances de chacune des résolutions aux transformations affines, en fonction de ces deux seuils (pour une valeur F1 maximale, ou supérieur à 95 % des distracteurs), sont présentées dans la Figure III-34. En dehors des échelles les plus basses (8 et 10 particulièrement), aux performances générales très faibles[1], on observe une tendance à une baisse de la tolérance lorsque la résolution augmente, résultat d'un traitement spatial plus fin, donc plus sensible aux changements locaux d'apparence.

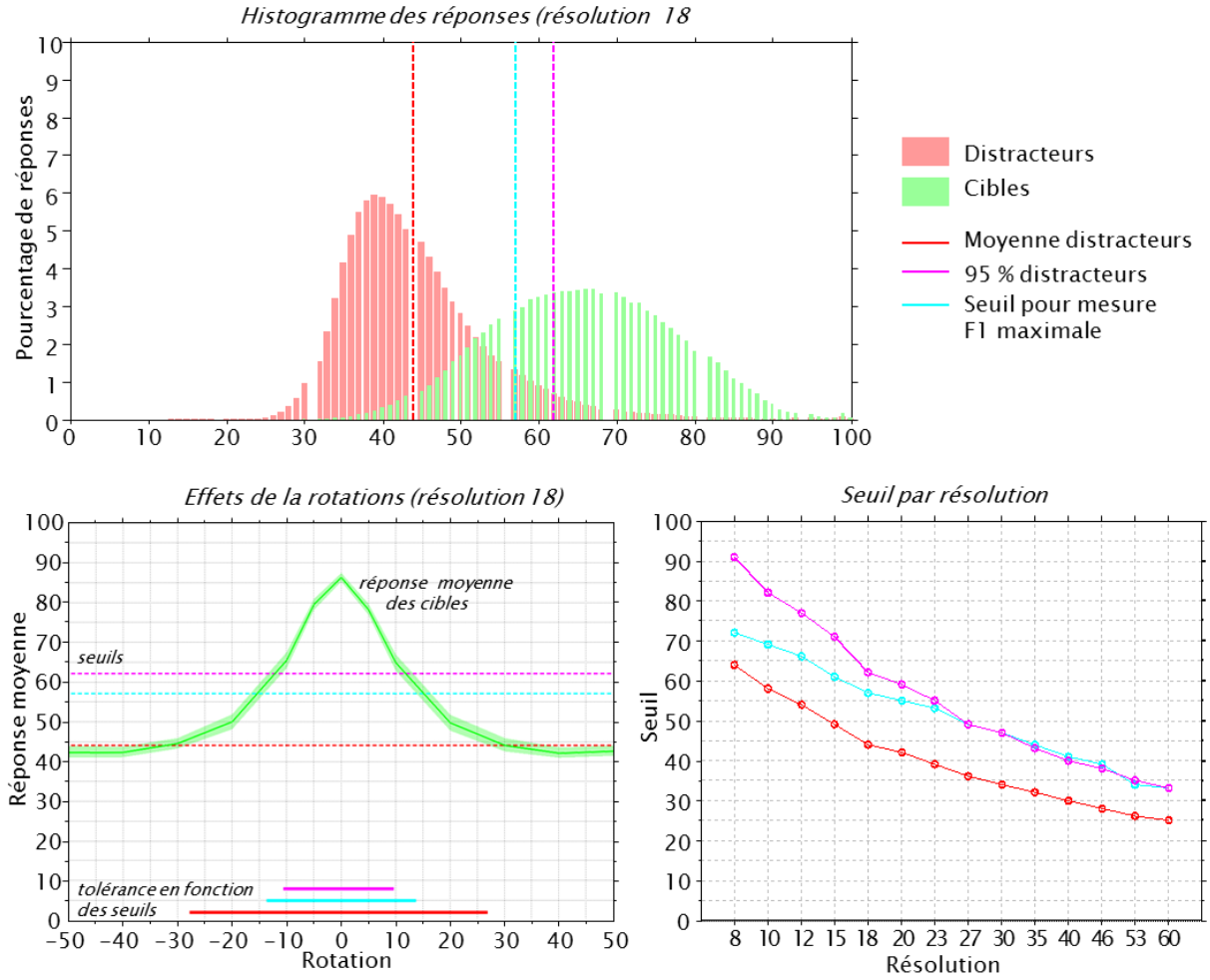


Figure III-33 Seuils utilisés pour l'évaluation de la tolérance aux transformations.

[1] Dues comme nous l'avions vu à une quantité d'information capturée trop faible.

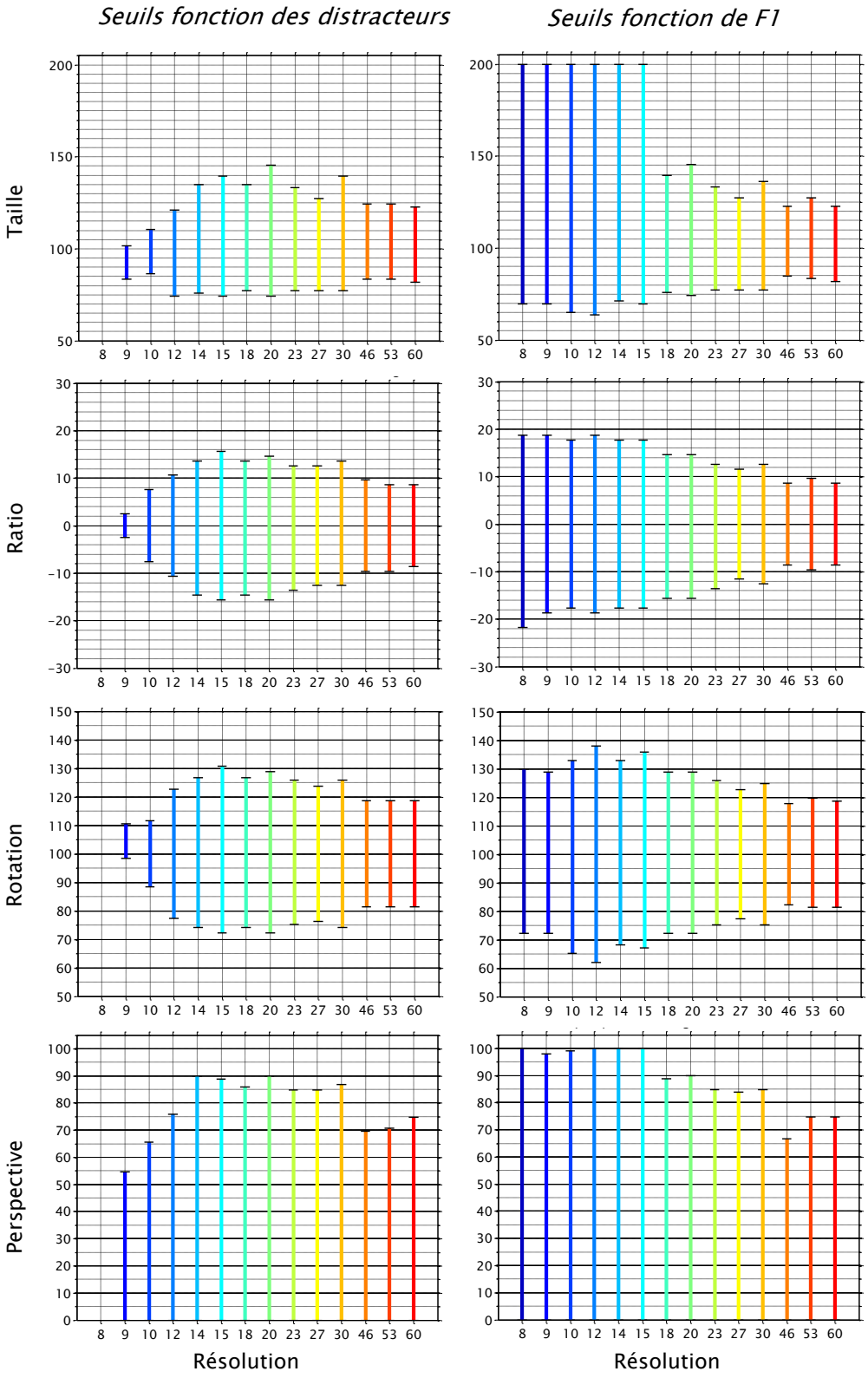


Figure III-34 Tolérance aux différentes transformations affines en fonction du choix des seuils à appliquer pour chacune des résolutions

3.3.5 Architecture finale

Tel que soulevé dans [Würtz and Lourens, 1997] les hautes échelles, si elles permettent de localiser de façon très précise les coins et arêtes, sont très sensibles aux changements locaux (bruit haute fréquence), et par conséquent peu robustes. A l'inverse les faibles résolutions sont plus stables mais imprécises concernant la localisation exacte, et de pouvoir discriminant inférieur. L'architecture MultiRes, illustrée dans la Figure III-35, vise à combiner ces différents avantages dans une approche *coarse to fine*, consistant à raffiner l'information dans une cascade de détections de complexités croissantes.

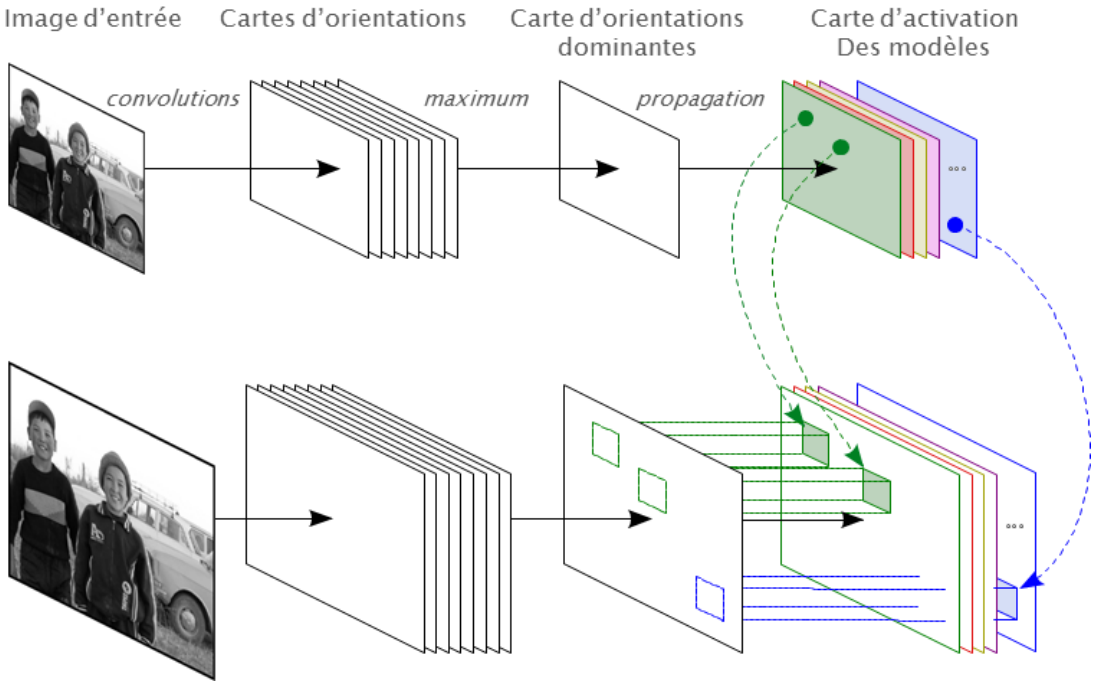


Figure III-35 Architecture Spikenet MultiRes

Comme nous venons de le voir, les basses résolutions permettent un gain de temps très conséquent. Elles souffrent néanmoins de taux de reconnaissance inférieurs aux échelles plus hautes, causés par une quantité de signal moins importante entraînant de nombreuses fausses détections. Nous proposons par conséquent une architecture en cascade. Une première passe à basse résolution, avec un seuil libéral, permettra de sélectionner les modèles potentiels et les régions candidates pour un traitement plus fin. Il est important d'utiliser ici des seuils relativement bas, car étant donnée la sensibilité plus faible de ces échelles, l'utilisation de seuils offrant une précision élevée[1], telle qu'elle pourrait être attendue en sortie du système, aurait pour conséquence de réduire fortement le

[1] C'est-à-dire peu de fausses alarmes. Par exemple, que parmi les détections, plus de 90% correspondent à la présence effective de la cible.

rappel et donc d'exclure des régions candidates. La deuxième passe, à une plus haute résolution, permettra ensuite d'effectuer un filtrage plus strict, par des seuils plus conservateurs.

Dans le projet d'utiliser cette architecture pour l'aide aux non-voyants, la rapidité des traitements est un facteur crucial, afin de garantir une utilisation temps-réel avec un maximum de modèles testés. Le choix des résolutions devra donc refléter ces aspects, la plus basse devant être suffisamment discriminante pour limiter le nombre de régions et de modèles candidats, tout en restant assez faible pour obtenir des gains de temps significatifs. La plus haute devant quant à elle offrir les meilleures performances générales en termes de précision et de rappel (comme nous l'avons vu précédemment, l'aire sous courbe ROC diminue au-delà d'une certaine taille, donc la résolution offrant les meilleurs résultats de classification ne correspond pas nécessairement à la plus haute).

Cependant nous ne nous intéressons pas seulement à gagner en rapidité pour des performances équivalentes, mais souhaiterions également augmenter la robustesse et la tolérance de l'algorithme monocouche originel. Ce gain passe par deux éléments. D'une part, la combinaison des réponses aux différents étages. Le score final pour la détection d'un modèle à une taille et position données réunira donc les votes des neurones aux différentes échelles, bénéficiant ainsi de la meilleure tolérance aux transformations affines des basses résolutions, que nous avons montrée plus tôt. Deuxièmement, il passe par le choix de résolutions aux réponses à la fois corrélées dans leurs hits (nécessité d'un accord des différentes résolutions lorsque la cible est présente), et décorrélées dans leurs fausses alarmes. Chaque modèle appris possède en effet deux constituantes, une certaine quantité de signal, relatif au stimulus encodé, ainsi qu'une part de bruit. Aux très basses résolutions par exemple, cette quantité de signal, trop faible, explique les performances médiocres. Si l'on considère deux échelles voisines, le signal capturé par chacune, lié au motif appris, sera en grande partie redondant (voir comme illustration les figures de la section 3.1.2), alors que le bruit est relativement indépendant à chaque échelle. En conséquence, la combinaison des réponses de ces deux résolutions risquera donc de n'apporter que très peu d'amélioration dans leurs performances de classification respectives (voir même potentiellement leur dégradation), le bruit étant multiplié, et le gain en signal très faible.

Grâce à l'architecture en cascade, la décorrélation des fausses alarmes permet de diminuer le seuil de la deuxième passe (haute résolution), sans augmenter son taux de fausses alarmes. En effet, une partie importante des régions où elle aurait à tort détecté une cible pourra être précédemment filtrée par la première passe. De façon réciproque, plus importante sera la décorrélation, plus faibles seront les chances que les zones où la cible est absente provoquent des détections dans la couche suivante, bien qu'elles aient été sélectionnées à basse résolution. Cette diminution des seuils d'activation, n'entrainant pas

d'augmentation des taux de faux positifs, permet ainsi des gains en termes de rappel et de précision, et donc en conclusion une amélioration générale des performances de l'algorithme.

En nous basant sur les différentes analyses des courbes ROC, des temps de traitement, tolérances aux transformations, ainsi qu'aux corrélations entre résolutions, nous avons finalement retenu les résolutions 18 et 27[1] dans notre modèle final, qui semblent offrir le meilleur compromis entre les différents facteurs considérés. L'apprentissage est, comme expliqué plus tôt, réalisé indépendamment à chaque échelle, en conservant un pourcentage de poids par rapport au nombre de pixels égal à celui du noyau classique (environ un tiers). Les différents modèles appris à chaque résolution sont conservés au sein d'un méta-modèle, qui sera utilisé dans la phase de reconnaissance. Lors de celle-ci, une première passe est réalisée en utilisant les modèles à basse résolution sur une ou plusieurs versions sous-échantillonnées de l'image testée, en fonction des paramètres de taille minimum et maximum[2] fixés. Si ceux-ci s'activent au-delà de leur seuil, ils seront alors recherchés dans la deuxième passe, grâce aux modèles haute-résolution associés, aux tailles et positions correspondant aux hits déclenchés. Si ces derniers dépassent à leur tour leur seuil d'activation, une détection sera levée, aux positions de la deuxième passe (plus précise) et au score combinant les réponses de chacune des résolutions (elles ne sont combinées que s'il s'agit du même modèle, à la même taille et position). Les gains de cette nouvelle architecture par rapport au noyau Spikenet classique sont illustrés dans la Figure III-36. Elle permet effectivement de gagner en tolérance aux transformations, mais en même temps en précision, grâce aux différents mécanismes que nous avons décrits. Les performances générales en termes d'aire sous courbes ROC se voient par conséquent améliorées, non seulement par rapport à l'algorithme Spikenet original (res 30), mais également par rapport aux deux résolutions qu'elle utilise prises indépendamment (res 18 et 27). Ainsi, sur la base de test de difficulté élevée, l'aire sous courbe ROC moyenne du noyau MultiRes est de 0.837 (intervalle de confiance à 95% : IC = [0.825 – 0.845]), alors qu'elle est de 0.806 (IC = [0.795 – 0.815]) pour le noyau classique. Sur la deuxième base de test, de difficulté moindre, elles sont respectivement de 0.990 (IC = [0.984 – 0.992]) et 0.977 (IC = [0.974 – 0.979]).

[1] Les modèles présentent donc une d'augmentation de taille de 150 % dans la seconde couche.
[2] S'ils sont à leur valeur par défaut (100 et 100), le modèle n'est recherché qu'à sa taille originale, ainsi que nous l'avons expliqué dans la sous-section précédente. Le pourcentage de redimensionnement de l'image testée correspondra donc au ratio entre la taille de la zone d'apprentissage (une région de 216 par 216 px par exemple), et la résolution du modèle (18 x 18 pour la première couche, entraînant par conséquent une division par 12 de la taille de l'image). Si l'on augmente l'intervalle de ce paramètre (par exemple 50-200), l'objet sera recherché sur plusieurs redimensionnements de l'image, couvrant la plage d'échelles fixée (la taille de l'image sera donc divisée par différents facteurs répartis entre 6 et 24, afin que l'objet puisse être trouvé à des tailles comprises entre 108 et 432 px).

Concernant les temps de traitement, les gains obtenus varient selon les images et les modèles testés. En effet, l'ensemble des modèles à rechercher est d'abord testé à basse résolution. Pour celle que nous avons choisie, 18x18, cette première passe est donc environ 8 fois plus rapide que le temps total de l'algorithme classique (à 30x30), tel que nous l'avons observé dans la section III.3.3.3. A à ce gain maximum, il faut néanmoins soustraire le temps de traitement nécessaire à la recherche des modèles sélectionnés dans la seconde passe. En fonction de leur seuil, de leurs caractéristiques (qu'il s'agisse d'un modèle fortement discriminant, ou un contraire assez pauvre, déclenchant de nombreuses détections), et des images testées, le nombre de modèles retenus et les zones candidates pourront donc varier, entraînant des temps de traitement plus ou moins long à haute-résolution. Dans notre corpus, le pourcentage de modèles correctement filtrés par la première passe était en moyenne de 93.57 % (avec un écart-type σ de 3.52 % sur les 850 images considérées). Le temps global nécessaire aux 2 passes était donc en moyenne 5,9 fois plus court que celui de l'algorithme mono-échelle[1] (σ = 0.8). Rappelons que dans le corpus que nous avons utilisé, les modèles appris sont tous extraits aléatoirement de photos prises de centre-ville, et donc potentiellement similaires. De même, plus de la moitié des images testées sont elles aussi des scènes urbaines, multipliant les chances de fausses alarmes. Dans un contexte plus varié, le nombre de modèles filtrés par la première passe à basse résolution serait donc probablement plus élevé que celui que nous avons mesuré, résultant en des vitesses de traitement encore supérieures.

[1] Les temps de traitements sont proportionnels au carré de la résolution et au nombre de modèles testés. Les deux étapes de l'algorithme étant consécutives, leurs temps respectifs sont à additionner. Seuls les modèles retenus lors de la première phase étant testés à haute résolution (environ 6 %), le gain de temps total correspond donc au facteur $30^4/(18^4 + 27^4 \times \frac{6}{100})$.

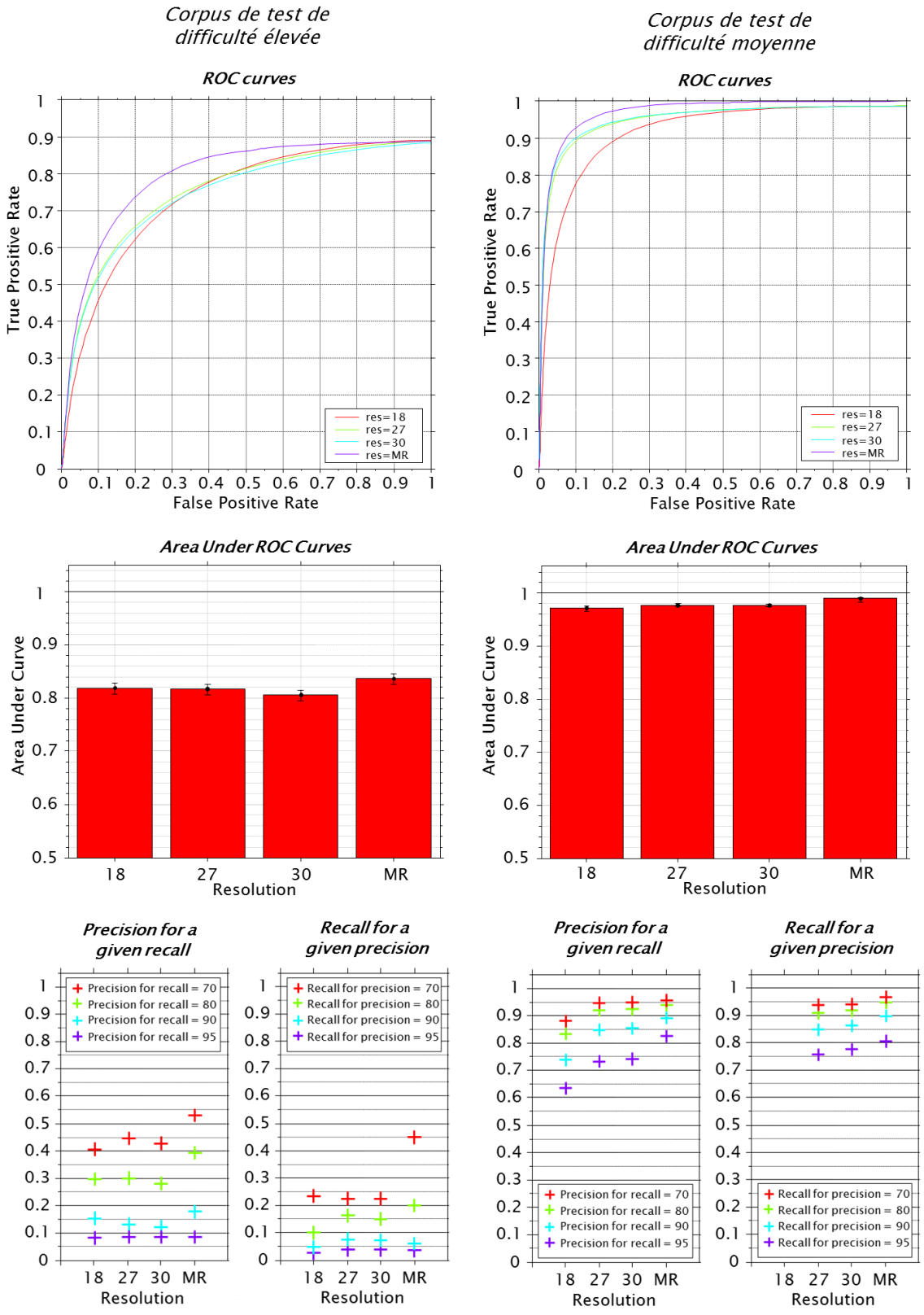


Figure III-36 Comparaison des performances de classification de l'architecture finale MultiRes (MR) avec les deux résolutions qu'elle utilise (18 et 27), ainsi qu'avec l'algorithme Spikenet original (res 30). Ces résultats (courbes ROC ; aires sous courbes ROC ; précision pour des valeurs de rappel fixées, et rappel pour des précisions données) sont fournis sur les deux bases de tests de différentes difficultés mentionnées précédemment.

4. Conclusion

Face aux performances des systèmes classiques de vision par ordinateur, ne parvenant toujours pas à se rapprocher des aptitudes humaines et animales en terme d'analyse d'une scène visuelle, et ce malgré l'explosion des ressources de calcul, nous observons depuis plusieurs années l'essor des approches bio-inspirées telles que les réseaux convolutionnels et le *deep learning*, dont les résultats sont de plus en plus prometteurs. Prenant comme référence l'organisation du système visuel, que nous détaillons en annexes, la plupart de ces algorithmes reposent sur des architectures relativement complexes, reflet des modèles traditionnels des sciences cognitives sur la reconnaissance d'objets. Ils peuvent par exemple intégrer la modélisation des mécanismes attentionnels (notamment par l'utilisation de cartes de saillance), de *grouping*, de segmentation, de *pooling*, et s'organisent généralement en un système comprenant de nombreuses couches de traitement successives (avec ou sans feedback).

L'étude, chez l'humain, des phénomènes de catégorisation rapide, a permis de proposer un autre modèle, plus simple mais extrêmement efficace, et rendant compte des temps de latence des réponses comportementales mesurés expérimentalement. En effet, dans les protocoles de choix saccadique, la plupart des sujets sont capables non seulement d'identifier, mais également de localiser de façon très précise une cible (telle qu'un visage) présentée dans le champ visuel périphérique, et ce en seulement 110 à 120 ms. Ces temps de réaction extrêmement courts (auxquels il faut de plus soustraire le déclenchement moteur de la saccade), combinés aux informations sur les latences de réponse des différentes aires corticales et sur la vitesse de conduction du signal nerveux dans le cerveau, semblent incompatibles avec les modèles de traitement hiérarchique de la voie ventrale généralement acceptés. Ceci ne remet évidemment pas en question ces modèles pour de nombreuses tâches de reconnaissance visuelle, mais suggère un autre type de traitement, plus court, lors de la détection de visages, d'animaux, ou d'autres stimuli écologiquement importants.

L'algorithme Spikenet est grandement inspiré de ces résultats et repose sur un petit nombre de couches de traitement reproduisant les traitements visuels précoces. Grâce au principe de son architecture et aux nombreuses optimisations apportées ces dix dernières années par la société du même nom, il offre comme atout majeur face aux autres systèmes une vitesse de traitement particulièrement courte. Cependant, s'il présente une invariance presque totale aux transformations photométriques de l'image, il est en revanche relativement assez sensible aux déformations résultant de transformations affines, et n'encode qu'une faible plage de fréquences spatiales. Ceci nous a amené dans cette thèse au

développement d'une nouvelle architecture, baptisée Spikenet MultiRes, qui reprend les forces de l'algorithme original en lui appliquant des traitements en cascade à différentes résolutions, permettant l'extraction d'informations plus riches sur les motifs visuels à apprendre. Son évaluation, détaillée dans la section III.3.3, a montré des gains en termes de tolérance aux transformations et de performances générales, mais également des vitesses de traitement encore plus courtes. Par conséquent, en permettant la recherche et la localisation d'un nombre potentiellement élevé de cibles en un minimum de temps, cet algorithme est particulièrement adapté à l'aide aux non-voyants, qui nécessite une interaction en temps réel avec l'environnement.

Plusieurs pistes permettant d'améliorer encore les performances de Spikenet MultiRes pourraient être suivies. Dans certains contextes de reconnaissance de formes, comme l'identification de caractères, l'invariance à la rotation ou à la symétrie n'est pas souhaitable (un 6 équivaudrait à un 9, un N à un Z,…). Pour d'autres, telle que la détection de visages, de piétons, de voitures, ou de bâtiments, contraindre la reconnaissance à une marge de rotations relativement faible permet souvent de réduire les risques de fausses alarmes, ces « objets » apparaissant généralement en position verticale. Cependant, pour une majorité d'objets (ou lorsque l'orientation de la caméra est inconnue), des mécanismes de reconnaissance insensibles aux déformations de l'image dans le plan sont gages de performances accrues. Si certaines méthodes décrites plus tôt permettent de conférer à Spikenet[1] la tolérance aux changements d'échelle et aux rotations, par la création dynamique de différents modèles, une solution plus complète et adaptative serait judicieuse. Ainsi, à partir d'une première segmentation de l'objet[2], il serait possible de générer lors de l'apprentissage un ensemble de transformations de l'image originale, telles que celles que nous avons appliquées pour constituer les corpus de tests. En apprenant un nouveau modèle pour chacune d'entre elles, nous pourrions alors constituer, à partir de leurs couvertures respectives, le sous-ensemble minimal assurant la détection de l'image dans l'ensemble des transformations choisies. C'est ce qui est habituellement fait de façon manuelle, en créant plusieurs modèles du même objet dans différentes images d'une vidéo, et en ajustant leurs seuils respectifs de manière à couvrir les changements d'apparence de l'objet tout en conservant un taux de fausses détections relativement bas. Par la procédure d'apprentissage semi-supervisée que nous venons d'évoquer, il serait possible, à partir d'une unique annotation, de constituer un lot de modèles optimal et de déterminer les paramètres de façon automatique au moyen d'une base de distracteurs permettant le réglage des seuils. Le temps d'apprentissage serait évidemment plus important, mais néanmoins bien plus court que dans le cas d'un apprentissage manuel. De plus, par la constitution d'un lot optimal de

[1] Aussi bien dans sa version classique que MultiRes.
[2] Consistant dans la plupart des cas à fixer la boite englobante de l'objet à apprendre dans l'image, et non à détourer celui-ci de façon précise.

modèles, cette méthode offrirait des gains de vitesse dans la phase de détection grâce au nombre réduit de sous-modèles. Pour terminer, il serait potentiellement intéressant d'évaluer plus en détail les performances de l'architecture proposée en augmentant le nombre de résolutions. Si quelques expérimentations utilisant une troisième passe encore plus fine ont donné des résultats inférieurs à ceux de la version proposée (comprenant deux échelles), il est toutefois possible d'imaginer les améliorer en choisissant la bonne combinaison de tailles et des paramètres adaptés (peut être en réintroduisant la compétition inter-échelle que nous avons évoqué, de sorte à ne conserver dans la résolution la plus haute que les saillances n'apparaissant pas également à des fréquences spatiales plus basses).

IV. Conclusion générale

Sommaire de section

1. Synthèse des contributions

Comme nous l'avons souligné dans l'introduction générale de cette thèse, le handicap visuel est un enjeu de société majeur. Touchant aujourd'hui près de 314 millions de personnes à travers le Monde, ses répercussions sur la vie quotidienne sont multiples [Thylefors et al., 1995; World Health Organization, 2005]. Il engendre en effet des incapacités dans un grand nombre de domaines allant de la communication à la mobilité, et entraîne donc une perte d'autonomie ainsi qu'une baisse significative de la qualité de vie [Salive et al., 1994; West et al., 2002]. Avec le vieillissement de la population, le nombre de déficients visuels va certainement continuer d'augmenter dans les prochaines années, jusqu'à doubler d'ici à 2030 selon plusieurs estimations [Foran et al., 2000; Frick and Foster, 2003; Taylor et al., 2005]. Il est donc nécessaire de prévenir et traiter les maladies entraînant des troubles visuels, lorsque cela est possible[1], mais il est tout aussi important d'améliorer la qualité de vie et l'autonomie des déficients visuels, objectifs que nous avons poursuivis dans cette thèse au travers du développement d'un système de suppléance basé sur la vision artificielle.

Les deux grandes approches holistiques pour compenser la perte ou l'absence de vision sont les systèmes de substitution sensorielle, restituant l'information visuelle par l'intermédiaire d'une autre modalité sensorielle (généralement l'audition ou le toucher), et les neuroprothèses visuelles[2]. De nombreux travaux ont effectivement montré qu'il était possible d'implanter des matrices d'électrodes dans le système nerveux afin de stimuler électriquement des relais sensoriels pour transmettre de l'information acquise par un récepteur artificiel. L'implant cochléaire en est un exemple fonctionnel. Plusieurs projets d'implantation des différents relais du système visuel (rétine, nerf optique, thalamus, cortex) sont ainsi en cours depuis la fin des années 1960. L'ensemble de ces travaux repose en outre sur une approche de type 'scoreboard'. Cela signifie que les informations acquises par des caméras sont reproduites par micro-stimulation à la surface du relais visuel implanté, sous la forme d'une matrice de points respectant la topographie de l'image. Ceci implique une énorme perte de résolution spatiale puisqu'une image de 640*480 pixels par exemple, n'est restituée que par quelques dizaines d'électrodes dans les interfaces actuelles (voir Figure IV-1). Il existe également une baisse de résolution temporelle puisqu'il est très compliqué de stimuler à la fréquence d'une caméra standard (autour de 30 Hz).

[1] A ce jour aucun traitement curatif ou préventif n'est par exemple connu pour la dégénérescence maculaire liée à l'âge, qui constitue la première cause de cécité dans les pays industrialisés.
[2] Se reporter à la section I.2.2 pour une description détaillée de ces méthodes.

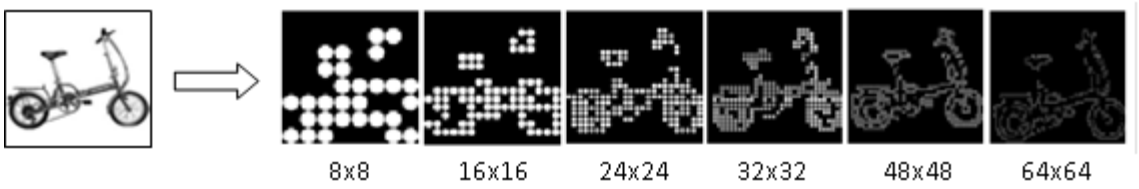


Figure IV-1 Simulation de carte de phosphènes en en fonction de la résolution d'un implant visuel (adapté de [Zhao et al. 2008]).

Ces limites sont de la même nature que celles des dispositifs de substitution sensorielle : la quantité d'informations visuelles nécessaire à l'interprétation d'une scène naturelle est bien trop importante par rapport à la résolution de l'interface de restitution (qu'elle soit auditive, tactile, ou par microstimulation). Leur efficacité reste donc très limitée. Ce constat nous a conduits à proposer une démarche alternative, consistant à intégrer des méthodes de vision artificielle permettant de prétraiter la scène visuelle afin de ne restituer au non-voyant que les informations extraites pertinentes.

Au sein d'un dispositif électronique de suppléance pour non-voyants, nous avons pu montrer que la détection en temps réel d'objets associée à la synthèse de sons spatialisés permettaient de réhabiliter des boucles visuomotrices qui rendent à nouveau possibles certaines fonctions visuelles comme la localisation et la préhension d'objets. La navigation étant une autre tâche critique pour les non-voyants, nous avons également incorporé au système des fonctions de guidage basées sur le positionnement par satellites et sur un système d'information géographique adapté. La faible précision de localisation du GPS, limitant l'usage des aides électroniques à l'orientation, nous a amenés à développer une nouvelle méthode de positionnement hybride, combinant les données satellites et inertielles à la reconnaissance de cibles visuelles géolocalisées. L'utilisation de la vision artificielle a permis d'obtenir des performances de localisation améliorées au cours de tests en conditions réelles, avec une erreur moyenne généralement inférieure à 5 mètres, et donc d'assurer le guidage et la navigation en temps-réel d'un piéton non-voyant.

Afin d'améliorer les performances du module de vision artificielle, constituant le cœur du système, nous avons développé et évalué un nouvel algorithme de reconnaissance de formes bio-inspiré, reposant sur la librairie Spikenet. Celle-ci utilise un codage de l'information visuelle par latence, et des représentations sous forme d'arêtes orientées, telles qu'observées dans le cortex visuel primaire. Par rapport à l'algorithme originel mono-échelle, cette architecture permet de capturer un spectre de fréquences spatiales plus large. Les traitements à faible résolution permettent ainsi d'améliorer la tolérance aux déformations de l'image, alors que les hautes fréquences spatiales, plus discriminantes, maintiennent une précision suffisamment élevée. De par son fonctionnement en plusieurs

passes successives, cette nouvelle architecture permet de plus de diminuer les temps de traitement grâce à une première couche rapide, filtrant les objets à rechercher dans la phase suivante à haute résolution, plus coûteuse en temps de calcul. Ces gains se sont avérés conséquents, multipliant par près de six la vitesse de l'algorithme.

Notons enfin que si ces travaux s'inscrivent dans le contexte de l'aide aux non-voyants, les résultats obtenus pourraient aussi être utilisés dans d'autres domaines d'application très différents. Ainsi, la méthode de positionnement que nous avons proposée, qui repose sur la détection d'amers visuels géolocalisés, pourrait être applicable à la robotique, au guidage de véhicules autonomes ou à l'amélioration de la précision des GPS piétons ou automobiles en cas de dégradation des signaux satellites. Elle permettrait aussi de compenser l'absence complète de signal GPS advenant par exemple à l'intérieur de bâtiments. L'algorithme de vision artificielle que nous avons développé, par sa robustesse et sa rapidité, serait quant à lui utilisable dans un grand nombre d'autres tâches nécessitant la reconnaissance de formes ou d'objets.

Pour terminer cette thèse, nous proposerons dans les sections suivantes une discussion de différentes questions liées à la vision artificielle et au système Navig (ou d'un point de vue plus général à l'assistance aux non-voyants), en proposant plusieurs perspectives aux travaux réalisés au cours de cette thèse. Nous discuterons ainsi de la mise en place des boucles sensorimotrices, des évolutions possibles de notre dispositif en termes de fonctionnalités et d'utilisabilité, de l'apprentissage automatisé des cibles visuelles par le système, ainsi que son application aux neuroprothèses.

2. Boucle sensorimotrice

Le système Navig, par l'association en temps réel de la détection d'une cible et de la restitution de sa position par des sons spatialisés, a permis la restauration d'une des fonctions majeure du système visuel, à savoir la localisation et la saisie d'objets. Une nouvelle boucle sensorimotrice est donc créée grâce au système de suppléance, couplant les mouvements de la tête et du corps avec la perception de sons provenant d'une source dans l'espace.

De nombreuses études ont démontré que l'action directe du sujet était nécessaire lors de l'utilisation de systèmes de substitution sensorielle [Arno et al., 2001b; Auvray and Myin, 2009; Bach-y-Rita, 1983]. Ainsi, lorsque la caméra permettant de capter la scène est fixe ou manipulée par une autre personne, les performances de reconnaissance ou de localisation sont extrêmement faibles, et les sujets rapportent l'expérience en termes de sensations ressenties dans la modalité de stimulation, au niveau de la peau par exemple pour les dispositifs visuotactiles. En revanche, le contrôle des mouvements de la caméra par l'utilisateur, qu'elle soit tenue à la main, ou fixée sur la tête, entraîne une extériorisation des percepts. La perception se différencie alors de la sensation. Cette différenciation est également supportée par le fait qu'une fois familiarisée au dispositif, la position de la matrice de stimulation peut être changée, tout comme celle de la caméra, sans affecter les performances et sans nécessité de réapprentissage [Bach-y-Rita, 2002], les objets de l'environnement étant alors perçus dans l'espace et non plus comme des sensations tactiles.

L'accès à l'information dans ce type de systèmes ne peut donc pas se faire par une exposition passive, mais nécessite au contraire une exploration de l'environnement par l'action. Cette hypothèse rejoint la théorie écologique de Gibson [Gibson, 1986, 1966] selon laquelle la perception est essentiellement active, et consiste en l'extraction d'invariants sensorimoteurs. Dans la même lignée, le modèle proposé par O'Regan et Noë définit lui aussi la vision comme une activité exploratrice de l'environnement qui permet l'apprentissage de contingences entre les mouvements de l'organisme et les sensations résultantes [Noë and O'Regan, 2002; O'Regan and Noë, 2001]. Différentes autres approches soulignent également le rôle de l'action dans la perception, comme l'approche enactive de la perception [Varela, 1992], postulant que les éléments perceptuels résultent du couplage entre l'environnement et l'organisme, ou encore l'*embodiment* proposé par Ballard [Ballard, 1991; Ballard et al., 1997].

Un système de suppléance, dès lors qu'il est contrôlé par l'individu, s'apparente donc à un organe sensoriel. Il peut en effet être l'objet d'actions, de mouvements, et permet

d'accéder aux propriétés de l'environnement. Les utilisateurs doivent donc apprendre à extraire les relations invariantes entre leurs actions et la stimulation sensorielle résultante, et la compréhension de ce couplage permet de développer, progressivement, un nouvel espace perceptif ouvert par l'outil. Différentes étapes de cet apprentissage sont proposées dans [Auvray, 2004] :

> *La première étape, l'étape du contact, implique l'apprentissage des régularités sensorimotrices nécessaires à stabiliser et à maintenir le contact perceptif avec le stimulus. Les utilisateurs doivent ainsi extraire les régularités entre les actions effectuées dans l'espace organique et les stimulations sensorielles résultantes.*
>
> *La seconde étape, l'attribution distale, est la compréhension, par les utilisateurs, que l'origine de leurs sensations organiques est due à la rencontre avec un objet provenant de l'espace perceptif ouvert par l'outil. Les utilisateurs doivent en premier lieu comprendre l'existence d'un espace perceptif nouveau. Puis ils doivent comprendre que les actions qu'ils effectuent dans leur espace organique leur permettent de déplacer des points d'actions (points de vue et / ou points d'inscription) dans l'espace distal. Et enfin, que les variations dans les stimulations sensorielles correspondent à des variations de points de vue sur des objets de l'espace perceptif distal.*
>
> *La maîtrise de l'espace distal est l'étape au cours de laquelle les utilisateurs apprennent à contrôler les nouvelles régularités sensorimotrices introduites par l'outil. Ils apprennent à faire varier les points de vue et les points d'inscriptions distaux. Ils parviennent ainsi à modifier l'espace perceptif distal et à localiser objets et événements de cet espace perceptif distal relativement à un point de vue appartenant à cet espace.*
>
> *La quatrième étape, la localisation distale, est l'impression d'être dans l'espace perceptif ouvert par l'outil. Cette étape implique, grâce à l'apprentissage, une automatisation du nouveau couplage sensorimoteur. Une fois que les utilisateurs parviennent à avoir un accès direct aux effets de leurs actions, sans avoir à réfléchir sur le maniement de l'outil ou sur le code utilisé, ils peuvent se sentir entièrement là où ils agissent (point de vue et point d'inscription), c'est-à-dire dans l'espace perceptif distal.*

Le système Navig, et plus particulièrement la fonction de localisation d'objets que nous avons proposée, offre ainsi toutes les caractéristiques pour la constitution d'un espace de perception distal suppléant à l'absence ou la détérioration de la vision chez ses utilisateurs. De plus, contrairement aux systèmes de substitution sensorielle auditifs

classiques (convertissant l'image selon un système de codage complexe), ou tactiles, l'apprentissage requis est ici extrêmement restreint. Les expérimentations avec des utilisateurs aveugles ou aux yeux bandés ont en effet démontré une utilisation presque immédiate du dispositif dans des tâches de guidage ou de saisie d'objets [Parseihian, 2012]. La rapidité de familiarisation à ce type d'interface sonore avait également pu être observée par les créateurs du Virtual Acoustic Space, un système à l'architecture similaire, qui retranscrit des cartes de profondeurs au moyen de sons spatialisés [Gonzalez-Mora et al., 2006, 1999]. Ce constat s'explique par le fait que les sons spatialisés reproduisent la perception de sources sonores réelles, et que les caméras, montées sur un casque ou des lunettes, se trouvent donc dans le même référentiel que la tête du sujet. Il n'y a par conséquent aucun apprentissage nécessaire des relations entre les mouvements de l'utilisateur et la stimulation résultante[1], celle-ci étant similaire aux conditions habituelles d'écoute et de localisation de sons.

[1] A l'exception de la familiarisation avec des HTRF génériques (les fonctions de transfert utilisées pour la synthèse binaurale), sujet discuté dans [Parseihian, 2012].

3. Convergence de fonctions visuelles

Les deux fonctionnalités majeures apportées par le dispositif Navig sont la reconnaissance de cibles visuelles pour le guidage ou la préhension, et la navigation vers une destination. Ces deux tâches comptent parmi les plus utiles et en même temps les plus délicates pour la population non-voyante. Néanmoins il existe un grand nombre d'autres problèmes rencontrés par les aveugles dans leur vie quotidienne qui pourraient être résolus par des solutions basées sur la vision artificielle.

La reconnaissance des billets de banque est un exemple de ces difficultés du quotidien. S'il existe quelques dispositifs à destination des non-voyants permettant de reconnaître les monnaies comme le Note Teller commercialisé par Brytech[1], le logiciel knfbReader[2], ou encore les systèmes proposés dans [Hasanuzzaman et al., 2011; Liu, 2008], ceux-ci montrent des performances insuffisantes (en temps nécessaire à l'identification d'un billet et/ou en taux de reconnaissance), ou un prix trop élevé. En intégrant un module dédié à cette tâche au prototype Navig, reposant également sur le moteur Spikenet, Parlouar et al. ont pu montrer l'efficacité d'une telle solution, avec une précision de 100% et un temps moyen autour de 10 secondes, soit 3 fois plus rapide que les systèmes existants [Parlouar et al., 2009].

De la même façon, plutôt que de développer de multiples systèmes ou dispositifs aux applications très spécifiques, dont le nombre, le prix, et l'encombrement rendraient l'usage difficile, nous pensons qu'un système unique et souple, comme celui que nous proposons avec Navig, pourrait permettre de résoudre un grand nombre de problème nécessitant une analyse visuelle. Déjà équipé du matériel et des composants logiciels nécessaires (stéréovision, reconnaissance de formes, synthèse vocale, sons spatialisés, etc.), il serait ainsi aisé et peu coûteux de développer d'autres extensions similaires à l'identification de billets. Parmi celles-ci nous pourrions par exemple inclure des fonctionnalités de reconnaissance de caractères (OCR[3]) génériques telles que celles réalisées dans [Chen and Yuille, 2011; Mattar et al., 2005; Silapachote et al., 2005], permettant entre autres la lecture de documents, de panneaux, d'enseignes. L'OCR permettrait également des applications plus spécifiques, comme l'identification des CDs, boites de conserves, et autres objets du quotidien. Il pourrait également être intéressant d'intégrer des options de détection de visages ou de personnes, tel que proposé dans [Hub et al., 2006a; Krishna et al., 2008,

[1] http://www.brytech.com/noteteller/
[2] http://www.knfbreader.com/
[3] *Optical Character Recognition*

2005; Ribeiro et al., 2012][1], ou encore de reconnaissance de code-barres permettant d'identifier une grande partie des produits de consommation [Kutiyanawala and Kulyukin, 2010; Tekin and Coughlan, 2010]. Le nombre d'applications est donc large, et la conception modulaire du système Navig permettrait dans l'avenir de proposer un outil complet à même d'aider les non-voyants dans de nombreuses situations.

[1] Le système présenté dans [Ribeiro et al., 2012] permet non seulement de détecter des visages (signalés par un son 3D) mais également de les identifier (lecture spatialisée du prénom à sa position dans le monde réel)

4. Ergonomie

Du point de vue de l'ergonomie, de nombreux aspects du système Navig sont encore perfectibles. D'un point de vue physique, de toute évidence, de gros efforts restent à faire pour disposer d'un équipement léger et utilisable au quotidien. Les prototypes ont été fabriqués artisanalement afin de pouvoir réaliser des tests fonctionnels et de démontrer la preuve de concept. Le dernier dispositif comprenait donc, comme nous l'avons vu, un casque sur lequel étaient montées des caméras stéréoscopiques BumbleBee, relativement lourdes, ainsi qu'une centrale inertielle. L'utilisateur portait également un boîtier GPS, un ensemble casque et micro, ainsi qu'un dernier capteur à la hanche, tous ces éléments étant reliés à un ordinateur portable transporté dans un sac à dos.

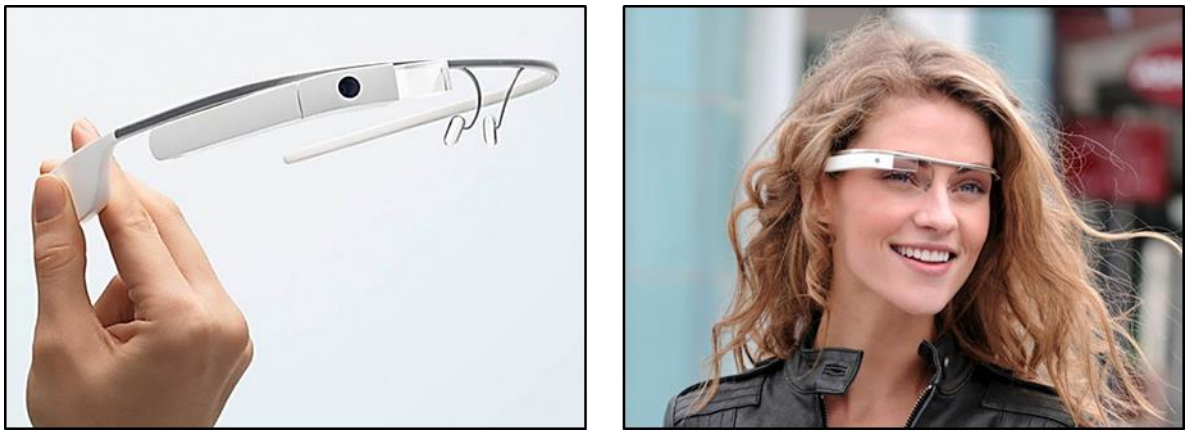


Figure IV-2 Google Glass

Avec l'apparition de produits tels que les Google Glass (voir Figure IV-2), Il semble néanmoins réaliste d'envisager la faisabilité d'un prototype compact intégrant les différents composants requis de l'architecture Navig. Ces Google Glass parviennent en effet à réunir dans une simple paire de lunettes, une caméra 5 millions de pixels, des modules Wi-Fi, GPS, et Bluetooth, des écouteurs à conduction osseuse, un micro, un capteur infrarouge, des gyroscopes, accéléromètres et magnétomètres 3 axes, ainsi qu'un système embarqué sur la base d'un processeur ARM. En attendant leur mise sur le marché prochaine, nous avons fait appel à Cambridge Research pour la réalisation d'un nouveau prototype, basé lui aussi sur des lunettes qui seront notamment équipées de deux petites caméras. Celui-ci, toujours en cours de réalisation, permettra des expérimentations avec des non-voyants en situation naturelle plus aisées qu'avec le matériel actuel. En plus des aspects de commodité liés au poids et à l'encombrement, le côté cosmétique joue également un rôle très important pour qu'un dispositif soit accepté et utilisé au quotidien. En effet, dans une enquête de Golledge,

de nombreux non-voyants ont même jugé ce critère comme plus important que sa capacité à les guider [Golledge et al., 2004]. Le côté compact et discret du dispositif aura donc aussi un rôle important à jouer dans les phases futures de développement.

En dehors des aspects techniques, les retours des premiers utilisateurs de Navig témoignent de l'importance de l'interface sonore. Même avec des modules très performants de positionnement, de reconnaissance d'objets, ou de cartographie, si les instructions et les sons fournis ne sont pas adaptés, l'usage du système restera difficile et peu agréable. Il est donc crucial de poursuivre la mise en place et l'amélioration de l'IHM afin que celle-ci réponde au mieux aux attentes des utilisateurs.

5. Apprentissage

Une problématique importante, peu abordée dans ce manuscrit, est la question de l'apprentissage. La mise en place et l'évaluation de la nouvelle architecture MultiRes a été réalisée en faisant apprendre au système des régions aléatoires d'images variées, celles-ci ne correspondant donc pas nécessairement à un objet précis ou à un élément à contenu sémantique, mais constituant simplement un motif visuel. L'utilisation de l'algorithme dans un contexte d'application précis soulève par conséquent la question de la création des modèles. Celle-ci passe habituellement par la sélection d'une zone au sein d'une image. Elle peut être également réalisée au moyen d'outils informatiques incorporant une interface graphique pour délimiter aisément la zone d'intérêt (comme le logiciel ModelBuilder intégré à la suite logicielle Spikenet), ou directement par des appels aux fonctions de la librairie de développement, qui fournissent l'image à apprendre et les coordonnées associées (l'ensemble de celle-ci, ou bien une sous-partie).

Comme nous l'avions mentionné en conclusion du troisième chapitre, afin de permettre la détection d'un objet à différentes tailles et orientations, il est possible de générer dynamiquement plusieurs sous-modèles couvrant ces transformations, à partir du modèle de référence. En revanche, si l'on souhaite pouvoir reconnaitre l'objet sous différents angles de vue, il est nécessaire de créer plusieurs modèles rendant compte des différentes apparences qui en résultent. Ceci est pour l'instant réalisé manuellement, autant pour l'algorithme Spikenet classique que pour sa version MultiRes, en apprenant un nouveau modèle de la zone d'intérêt chaque fois que ceux créés jusque-là ne parviennent pas à dépasser leur seuil d'activation (du fait de changements trop important du point de vue).

Plusieurs méthodes me semblent permettre de résoudre ce problème de façon automatique. La première, que nous avons partiellement implémentée, consiste en un créateur semi-supervisé qui reposesur le suivi de cibles. Il peut s'apparenter au *Co-Training* [Blum and Mitchell, 1998] ou au *Self-Training* [Agrawala, 1970], deux techniques permettant d'incorporer incrémentalement des exemplaires non annotés à un algorithme d'apprentissage, en sélectionnant ceux qui ont les plus hauts scores de prédiction. Dans notre contexte d'application, nous proposons de générer automatiquement des modèles assurant la reconnaissance d'un objet sous différentes poses par un suivi dans des images consécutives. Ainsi, à partir d'une unique segmentation manuelle dans une image quelconque d'une vidéo de l'objet (prise en variant l'angle de vue), il est possible de constituer un lot de modèles couvrant l'ensemble des changements d'apparence observés. Comme illustré dans la Figure IV-3, la première annotation permet la création d'un modèle initial. Celui-ci est ensuite recherché dans l'image suivante. Les coordonnées et l'échelle de

la détection au score maximum dans cette nouvelle image permettent alors l'apprentissage d'un nouveau modèle dans la zone où l'objet a été trouvé. Ce deuxième modèle est de la même façon propagé dans la troisième image[1], et le suivi est ainsi répété itérativement jusqu'à la fin de la vidéo. Cette phase de propagation permet donc de constituer une large collection de modèles de l'objet à apprendre (un pour chaque image où il est présent).

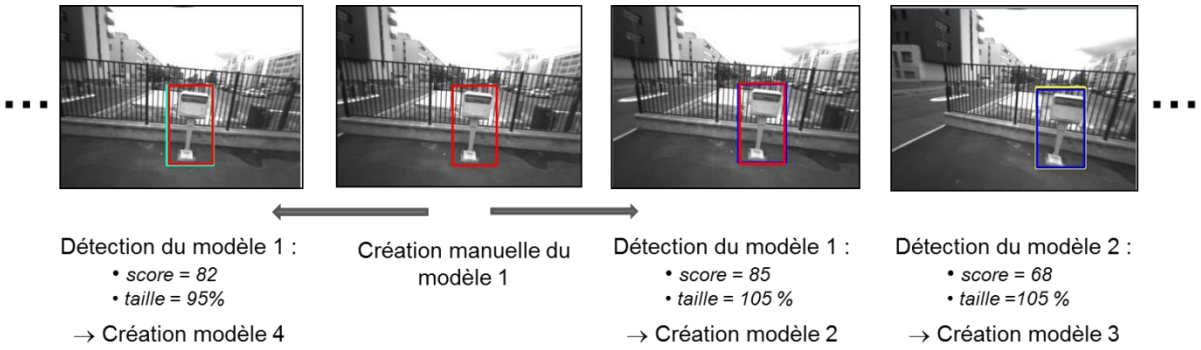


Figure IV-3 Propagation de la création de modèle par un suivi de proche en proche de l'objet dans les images de la vidéo

L'intégralité de ces modèles pourraient évidemment être utilisée pour la reconnaissance de l'objet, mais leur nombre rendrait cette solution trop coûteuse en temps de calcul. Nous avons donc mis en place une méthode de sélection d'un sous-lot optimal. Celle-ci passe par le test de chacun des modèles sur toutes les images de la vidéo en appliquant des seuils de détection relativement bas. A partir de leurs valeurs d'activation, et de la vérité terrain[2], il est possible de déterminer automatiquement le seuil d'activation qui garantit un taux admissible de fausses alarmes. Enfin, en utilisant ces seuils pour filtrer les détections précédemment effectuées, nous obtenons la couverture de chacun des modèles (voir Figure IV-4). Celles-ci nous permettent finalement de déterminer le sous-ensemble minimal de modèles maximisant la couverture sur l'ensemble des apparitions de l'objet dans la vidéo. Une première version de cette méthode de création de modèle semi-supervisée a pu être implémentée par un stagiaire en master, mais souffre encore de problèmes liés au suivi. La propagation de frame en frame entraîne en effet souvent une dérive progressive de la position estimée de la cible. Après un nombre important d'images, ces erreurs cumulées finissent par décentrer la zone supposée de l'objet, et les modèles alors créés ne sont donc plus représentatifs de l'objet à apprendre. Des améliorations du *tracking* sont donc

[1] Si l'image annotée manuellement n'est pas la première de la vidéo, ce processus de propagation est évidement à réaliser dans les deux directions, en suivant l'objet dans les parties précédant et suivant l'image de référence.
[2] La phase de propagation a enfin permis de déterminer les images de la vidéo où l'objet est présent, ainsi que sa position.

nécessaires pour disposer d'un créateur automatisé robuste, en utilisant par exemple un algorithme plus précis que Spikenet[1] afin d'assurer le suivi dans la phase de propagation, où en constituant un lot initial de modèles plus limité[2].

Scores de détections

Frames	1	2	3	4	5	7	8	9	10	11	12	13	14	15	16	17	18	19
Modèle1			45	61	75	67			34			39						46
Modèle2	54	51						62				87	91	98	88	74		

Couverture

Frames	1	2	3	4	5	7	8	9	10	11	12	13	14	15	16	17	18	19
Modèle1 – S=61			45	61	75	67			34			39						46
Modèle2 – S=87	54	51						62				87	91	98	88	74		

Figure IV-4 Calcul de la couverture de chacun des modèles. Les frames de la vidéo sont sur fond vert lorsque la cible est présente, rouge quand elle est absente. Les scores de détection de deux des modèles sont indiqués pour chaque image, de couleur verte s'il s'agit d'un vrai positif, rouge si c'est une fausse alarme. La couverture peut être calculée pour chacun en fonction du taux de fausses alarmes souhaité (un taux de 0.1 par exemple fixerait le seuil *S* du modèle 1 à 39 et celui du modèle 2 à 74, alors que pour un taux égal à 0, le seuil du modèle 1 augmenterait à 61). Les cellules en noir correspondent aux images couvertes par le modèle au seuil fixé.

Si elle ne garantit pas d'obtenir un lot de modèles optimal, une démarche extrêmement moins coûteuse[3], inspirée de l'*adaptative tracking* (voir par exemple [Jacquot et al., 2005; Javed et al., 2005; Kalal et al., 2009]), consiste à générer de nouveaux modèles chaque fois que le score d'activation d'un objet détecté diminue de façon sensible. L'apprentissage d'un visage vu de face n'entraînera par exemple pas de détection d'une vue de profil, les motifs visuels étant bien trop différents. En revanche, si le visage pivote lentement, et que l'on apprend un nouveau modèle pour chaque détection ayant un score en dessous d'une valeur donnée (mais supérieure au seuil d'activation du modèle), on obtiendra un jeu de modèles couvrant toutes les orientations, sans perte du suivi de la cible. Cette technique se rapproche de la méthode précédemment proposée, dans le sens où elle enrichit

[1] Les modèles Spikenet sont en effet appris à une résolution de 900 pixels (30 par 30 par exemple), donc les coordonnées de détection dans des images de taille importante peuvent présenter des erreurs de l'ordre de quelques pixels.
[2] Les légères imprécisions de localisation intervenant à chaque détection se cumulant, on réduirait la dérive observée en ne créant pas un modèle pour chacune des images mais à des intervalles plus espacés.
[3] Et suffisamment rapide pour être effectuée en temps réel.

progressivement les représentations de l'objet à partir des détections de modèles précédemment appris, mais elle est computationnellement beaucoup plus simple et pourrait ainsi permettre la création de modèles « à la volée » pour supporter le changement de pose des objets détectés. En utilisant des modèles « d'amorce » relativement génériques à des seuils libéraux, puis en générant dynamiquement des modèles plus spécifiques, elle permettrait également de prendre en compte les différences dans l'apparence d'objets d'une même catégorie. Pour reprendre l'exemple de la détection de visages, on observe en effet une certaine variabilité dans l'apparence des visages, et ce même à des orientations identiques (à cause de différences individuelles, mais également d'autres facteurs pouvant varier pour une même personne comme le port de lunettes, d'un chapeau ou le changement de conditions d'éclairage). Un lot de modèles prototypiques utilisant des seuils assez tolérants pourrait donc permettre une première détection d'un visage donné, déclenchant l'apprentissage de modèles spécialisés offrant une reconnaissance plus robuste et générant moins de fausses alarmes.

Enfin, dans certains cas particuliers où seul l'objet d'intérêt est en mouvement dans la scène visuelle[1], il suffit de segmenter les régions changeantes (grâce à des méthodes utilisant la soustraction de fond, le calcul de flots optiques, ou les différences entre trames successives [Richefeu, 2006]), afin d'obtenir la zone à apprendre [Murase and Nayar, 1995]. Cette solution pourrait s'avérer très utile à la problématique d'aide aux déficients visuels. En effet, l'ensemble des systèmes permettant la reconnaissance automatique d'objets nécessitent le concours d'une personne voyante (souvent même les développeurs du système) pour que de nouveaux objets soient ajoutés, ce qui constitue un verrou majeur à la généralisation de cet type de dispositifs et à une utilisation réelle dans la vie quotidienne. L'apprentissage implique en effet de prendre une ou plusieurs photos, et de segmenter manuellement l'objet dans celles-ci. Dans un dispositif comme Navig, où l'utilisateur est équipé de caméras montées sur un casque, il est possible d'imaginer une procédure d'ajout qui puisse être réalisée de façon autonome par un non-voyant. Elle consisterait à activer la fonction d'apprentissage, puis à tenir l'objet face à soi en le faisant pivoter doucement, tout en maintenant la tête relativement fixe[2], et finalement donner verbalement son nom ou sa description qui sera enregistrée conjointement au modèles appris.

[1] Plusieurs bases d'apprentissage ont par exemple été constituées en filmant des objets placés sur un support rotatif pivotant de façon régulière [Kim et al., 2007; Moreels and Perona, 2007].
[2] Même en cas de légers mouvements des caméras, les algorithmes de flots optiques permettent de différencier les changements globaux qui en résultent des mouvements locaux de l'objet.

Figure IV-5 Vues du Capitole de Toulouse fournies par Google Street View, Bing Maps StreetSide, Villes en 3D (Pages Jaunes) et Mappy UrbanDive

Le système Navig utilise par ailleurs une deuxième catégorie de cibles visuelles, les amers géolocalisés, qui permettent le raffinement du positionnement de l'utilisateur (déterminer ses coordonnées géodésiques à partir de la vision artificielle pour corriger l'imprécision du GPS). La création de ces Points Visuels (PV), tels que nous les avons baptisés, soulève d'importantes questions quant à une l'application possible de cette méthode à de nouveaux lieux. En effet, pour obtenir une localisation précise pouvant supporter la navigation piétonne d'un non-voyant, des PV doivent pouvoir être détectés dans une majorité des espaces extérieurs traversés par l'utilisateur[1]. Au cours des expérimentations du système Navig, les trajets se limitaient à des parcours de test prédéfinis. Des modèles Spikenet avaient donc été créés manuellement dans les endroits parcourus à partir de vidéos prises sur les lieux, tout en annotant dans le système d'information géographique leur position correspondante. Ce mode de création n'est évidemment pas transposable à une solution générique utilisable par un grand nombre d'utilisateurs. Nous proposons donc d'utiliser des services similaires à Google StreetView, collectant à partir de véhicules dédiés des images omnidirectionnelles ainsi que d'autres données telles que la position GPS, les réseaux wifi disponibles, ou encore des relevés 3D de l'environnement urbain. Ce type de projets tend à se démocratiser, et couvrent aujourd'hui la plupart des villes (du moins dans les pays développés). De nombreuses sociétés privées telles que Bing, Google, Mappy ou des services municipaux permettent maintenant l'accès à des images de la quasi-totalité des rues françaises (voir Figure IV-5). La commune d'agglomération du Grand Toulouse nous a par exemple fourni des relevés de la ville extrêmement précis grâce à l'utilisation de données lasers alignées aux photographies, et à

[1] En particulier dans les environnements urbains, où les conditions GPS sont les plus dégradées.

des systèmes de positionnement avancés. Accessibles via une librairie de développement ou par le logiciel présenté dans la Figure IV-6, ces données permettent notamment de récupérer les coordonnées GPS (longitude, latitude et altitude) de n'importe quel point dans une des images recueillies. Il est donc possible d'ajouter automatiquement dans le SIG la position de nouveaux PVs. Il est plus délicat en revanche de choisir, de façon non-supervisée des régions dans ces images pouvant constituer de « bons » amers visuels. Une solution consisterait par exemple à choisir aléatoirement[1] des zones candidates, puis à tester les modèles créés à partir de celles-ci sur d'autres images pour déterminer leur seuil d'utilisation et leurs performances de classification. Ainsi les modèles peu caractéristiques (appris sur une partie uniforme d'un mur par exemple) seront rejetés, alors que ceux reflétant des motifs facilement discriminable seront ajoutés au SIG avec leurs coordonnées associées. Evidemment, une telle méthode serait particulièrement coûteuse en temps de traitement, mais la vitesse n'est dans cette situation pas un facteur critique car ces calculs peuvent être effectués hors-ligne.

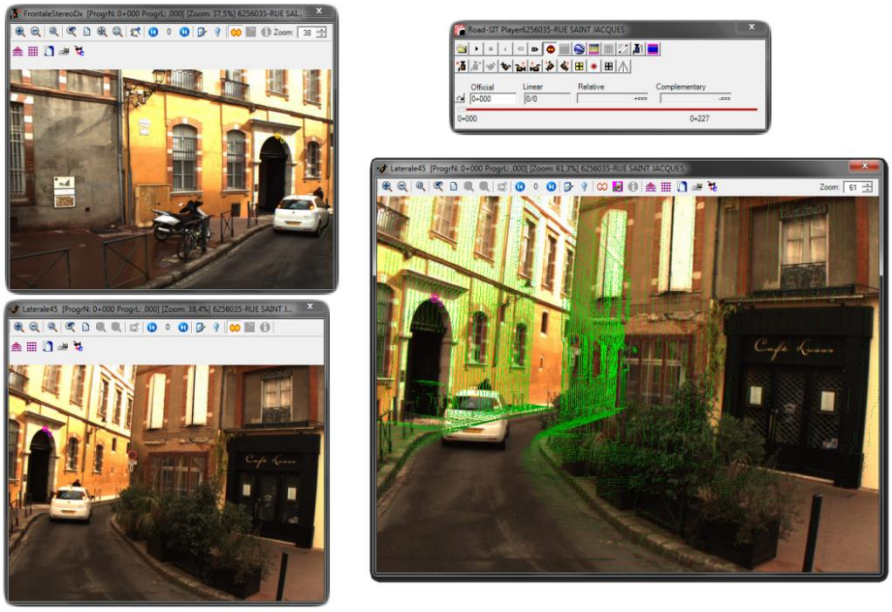


Figure IV-6 Logiciel de gestion des données visuelles, laser, et géodésiques recueillies par la société Siteco pour la communauté d'agglomération du Grand Toulouse

[1] Ou en sélectionnant préférentiellement des zone saillantes de l'image (où l'énergie locale est la plus haute par exemple), ayant plus de chance de contenir des caractéristiques.

6. Neuroprothèses

Les avancées du projet Navig et des travaux réalisés au cours de cette thèse s'inscrivent dans le cadre des systèmes d'aide aux déficients visuels par des dispositifs électroniques réhabilitant certaines fonctions spécifiques. Si notre système reposait sur une interface sonore, l'approche poursuivie, qui consiste à prétraiter la scène par des algorithmes de vision artificielle et à ne restituer que l'information visuelle haut-niveau pourrait être appliquée aux neuroprothèses, en transmettant les informations extraites par stimulation directe.

Il serait alors possible, même avec un nombre très limité de phosphènes, de restaurer certains comportements visuomoteurs grâce à la localisation d'objets. Jason Dowling déclarait à ce sujet en 2004 [Dowling et al., 2004]:

> *The use of image processing could enhance the effectiveness of visual prosthesis systems. We can use an information reduction approach to provide essential environmental information, and/or attempt to understand objects in the environment. Most existing visual prosthesis efforts are aimed at the information reduction level, which is concerned with the reduction or collapse of visual information. Edge detection is a useful method of encoding and describing information from an image in a more economical form, and involves identifying image contours where the brightness of the image changes abruptly. [...] A different approach involves attempting to understand components of the scene. This scene understanding level is concerned with identifying features and extracting information. The scene structure is still there to a degree, but it is idealised or reduced. An example application might be to identify a bus stop, fire hydrant or traffic light. It may also be useful to know the distance to the object (number of steps, or time at current walking speed). Due to the limited number of phosphenes that can be generated by current technology, it may be better to present a symbolic representation. For example a small part of the grid (perhaps 5x5) could be used for information on obstacle locations in the current environment. Auditory Information could also be provided in natural language, for example "A door is located forward to the right". A scene description mode could be useful.*

Malgré ce constat, dix ans plus tard presqu'aucun des groupes travaillant au développement d'implants visuels n'a suivi cette voie et n'utilise des traitements haut-niveaux de la scène afin d'en extraire des informations sémantiques. Certains travaux récents commencent toutefois à tirer parti des méthodes de vision artificielle [Barnes, 2013].

Plusieurs proposent par exemple des zooms interactifs dans certaines zones d'intérêt du champ visuel afin de permettre une analyse plus fine malgré le nombre limité de phosphènes : les régions où des visages sont détectés [He et al., 2012], ou celles dans la direction du regard[1] [Horne et al., 2012] (voir Figure IV-7). D'autres enrichissent l'information transmise au moyen de divers procédés tels que la détection du sol [Lui et al., 2012; McCarthy et al., 2011] (pour l'évitement d'obstacles et pour la navigation), le calcul de cartes de profondeurs[2] [Li et al., 2012; McCarthy and Barnes, 2012], ou des représentations symboliques [Lui et al., 2012], comme illustré dans la Figure IV-8.

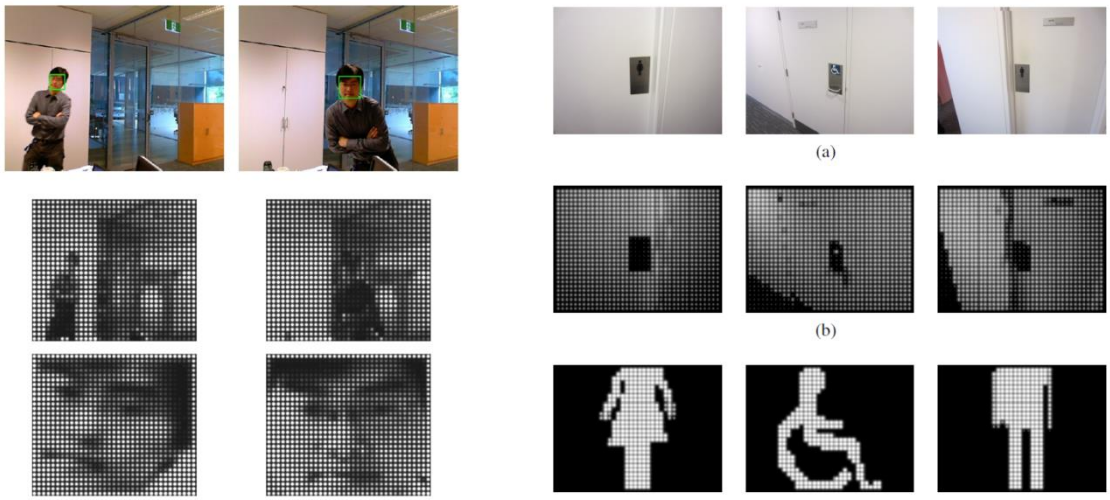


Figure IV-7 Zoom interactif dans des zones d'interêt de la scène. A gauche, la méthode proposée par [He et al., 2012], à droite celle de [Horne et al., 2012]. La première ligne correspond aux images acquises par les caméras, les suivantes à leur représentation sous forme de phosphènes sans ou avec zoom.

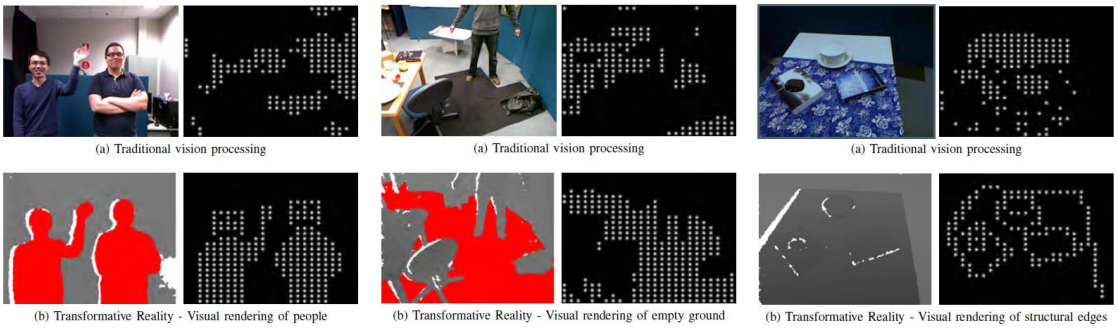


Figure IV-8 Représentations classiques (sur la première ligne) et augmentées grâce à la vision artificielle (sur la seconde) [Lui et al., 2012].

[1] Ou de l'orientation de la caméra si le système n'inclut pas de suivi des mouvements oculaires.
[2] Utiles lorsque des pixels proches de l'image ont des valeurs d'intensité voisines alors qu'ils correspondent à des points à des distances différentes (un objet au premier plan de couleur similaire à l'arrière-plan par exemple). Les contrastes de disparités seront donc renforcés lors de la stimulation pour que ces phosphènes puissent être différentiables.

Bien qu'elles soient encore rares (l'approche *scoreboard* restant la norme), ces solutions nous semblent très prometteuses et confortent l'approche que nous avons proposéz dans cette thèse. La vision artificielle apparaît en effet être la clé pour développer des systèmes d'aides aux non-voyants efficaces.

Différents travaux sont actuellement menés dans l'équipe Elipse de l'IRIT sur la simulation d'une neuroprothèse visuelle dans un casque de réalité virtuelle (voir Figure IV-9), afin de comparer le nombre, la position et la combinaison des phosphènes[1] nécessaires pour réaliser des tâches de localisation et de préhension d'objets, en utilisant une conversion de la scène de type *scoreboard* ou en utilisant des prétraitement par des algorithmes de vision artificielle [Denis et al., 2013, 2012]. Un projet d'implant a été initié en collaboration avec le CerCo , leLAAS (nanotechnologies) et le service de Médecine Physique et rééducation du CHU de Rangueil. Dans ce système, les positions des phosphènes perçus correspondent à l'emplacement dans le champ visuel des objets détectés au moyen du module de vision artificielle que j'ai développé au cours de cette thèse. Les améliorations des méthodes de détection, et les résultats que j'ai obtenus dans le cadre du projet Navig trouvent donc une application directe dans ces recherches visant à proposer un nouveau type d'implants visuels centré sur les aspects fonctionnels.

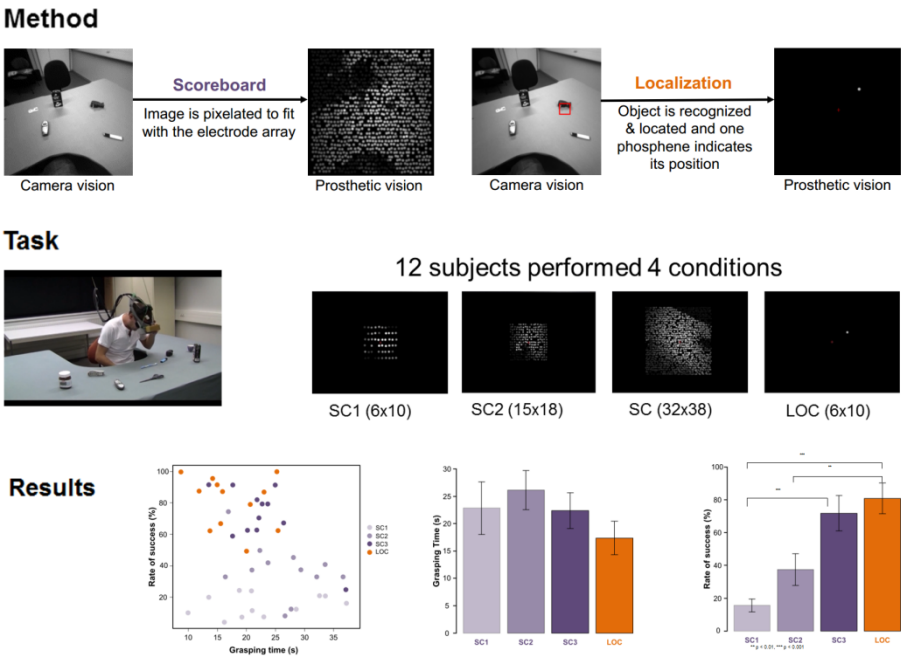


Figure IV-9 Comparaison entre l'approche scoreboard et l'augmentation sensorielle par vision artificielle dans une tâche de reconnaissance et de localisation d'objetq [Denis et al., 2012].

[1] Qui se traduisent par le nombre, l'emplacement et la spécification des électrodes à implanter.

Références

Adams, C.J., Beaton, R.J., 2000. An investigation of navigation processes in human locomotor behavior, in: Proceedings of the Human Factors and Ergonomics Society Annual Meeting. SAGE Publications, pp. 233–236.

Adjouadi, M., 1991. A man-machine vision interface for sensing the environment. J. Rehabil. Res. Dev. 29, 57–76.

Agrawala, A., 1970. Learning with a probabilistic teacher. IEEE Trans. Inf. Theory 16, 373–379.

Aguerrevere, D., Choudhury, M., Barreto, A., 2004. Portable 3D sound/sonar navigation system for blind individuals, in: The 2nd LACCEI Int. Latin Amer. Caribbean Conf. Eng. Technol. Miami, FL.

Ahuja, A.K., Dorn, J.D., Caspi, A., McMahon, M.J., Dagnelie, G., daCruz, L., Stanga, P., Humayun, M.S., Greenberg, R.J., Argus II Study Group, 2010. Blind subjects implanted with the Argus II retinal prosthesis are able to improve performance in a spatial-motor task. Br. J. Ophthalmol. 95, 539–543.

Aigrain, P., Zhang, H., Petkovic, D., 1996. Content-based representation and retrieval of visual media: A state-of-the-art review. Multimed. Tools Appl. 3, 179–202.

Aitken, S., Bower, T.G.R., 1983. Developmental Aspects of Sensory Substitution. Int. J. Neurosci. 19, 13–19.

Al-Khalifa, H.S., 2008. Utilizing qr code and mobile phones for blinds and visually impaired people, in: Computers Helping People with Special Needs. Springer, pp. 1065–1069.

Amedi, A., Stern, W.M., Camprodon, J.A., Bermpohl, F., Merabet, L., Rotman, S., Hemond, C., Meijer, P., Pascual-Leone, A., 2007. Shape conveyed by visual-to-auditory sensory substitution activates the lateral occipital complex. Nat. Neurosci. 10, 687–689.

Arikawa, M., Konomi, S., Ohnishi, K., 2007. Navitime: Supporting pedestrian navigation in the real world. Pervasive Comput. IEEE 6, 21–29.

Arno, P., Capelle, C., Wanet-Defalque, M.-C., Catalan-Ahumada, M., Veraart, C., 1999. Auditory coding of visual patterns for the blind. Perception 28, 1013–1030.

Arno, P., De Volder, A.G., Vanlierde, A., Wanet-Defalque, M.-C., Streel, E., Robert, A., Sanabria-Bohórquez, S., Veraart, C., 2001a. Occipital activation by pattern recognition in the early blind using auditory substitution for vision. Neuroimage 13, 632–645.

Arno, P., Vanlierde, A., Streel, E., Wanet-Defalque, M.-C., Sanabria-Bohorquez, S., Veraart, C., 2001b. Auditory substitution of vision: pattern recognition by the blind. Appl. Cogn. Psychol. 15, 509–519.

Atick, J.J., Redlich, A.N., 1990. Mathematical model of the simple cells in the visual cortex. Biol. Cybern. 63, 99–109.

Atick, J.J., Redlich, A.N., 1992. What does the retina know about natural scenes? Neural Comput. 4, 196–210.

Attebo, K., Mitchell, P., Smith, W., 1996. Visual Acuity and the Causes of Visual Loss in Australia: The Blue Mountains Eye Study. Ophthalmology 103, 357–364.

Auvray, M., 2004. Immersion et perception spatiale. L'exemple des dispositifs de substitution sensorielle. ECOLE DES HAUTES ETUDES EN SCIENCES SOCIALES.

Auvray, M., Hanneton, S., Lenay, C., O'REGAN, K., 2005. There is something out there: distal attribution in sensory substitution, twenty years later. J. Integr. Neurosci. 4, 505–521.

Auvray, M., Hanneton, S., O Regan, J.K., 2007. Learning to perceive with a visuo-auditory substitution system: Localisation and object recognition withThe vOICe'. Percept.-Lond.- 36, 416.

Auvray, M., Myin, E., 2009. Perception With Compensatory Devices: From Sensory Substitution to Sensorimotor Extension. Cogn. Sci. 33, 1036–1058.

Bach-y-Rita, P., 1983. Tactile vision substitution: past and future. Int. J. Neurosci. 19, 29–36.

Bach-y-Rita, P., 2002. Sensory substitution and qualia. Vis. Mind 497–514.

Bach-y-Rita, P., Collins, C.C., Saunders, F.A., White, B., Scadden, L., 1969a. Vision substitution by tactile image projection.

Bach-y-Rita, P., Collins, C.C., White, B., Saunders, F.A., Scadden, L., Blomberg, R., 1969b. A tactile vision substitution system. Am. J. Optom. Arch. Am. Acad. Optom. 46.

Bach-y-Rita, P., Kaczmarek, K.A., Tyler, M.E., Garcia-Lara, J., 1998. Form perception with a 49-point electrotactile stimulus array on the tongue: A technical note. J. Rehabil. Res. Dev. 35, 427–430.

Bailey, T., Durrant-Whyte, H., 2006. Simultaneous localization and mapping (SLAM): Part II. Robot. Autom. Mag. IEEE 13, 108–117.

Bak, M., Girvin, J.P., Hambrecht, F.T., Kufta, C.V., Loeb, G.E., Schmidt, E.M., 1990. Visual sensations produced by intracortical microstimulation of the human occipital cortex. Med. Biol. Eng. Comput. 28, 257–259.

Balakrishnan, G., Sainarayanan, G., Nagarajan, R., Yaacob, S., 2007. Wearable Real-Time Stereo Vision for the Visually Impaired. Eng. Lett. 14.

Ballard, D.H., 1991. Animate vision. Artif. Intell. 48, 57–86.

Ballard, D.H., Hayhoe, M.M., Pook, P.K., Rao, R.P., 1997. Deictic codes for the embodiment of cognition. Behav. Brain Sci. 20, 723–742.

Bamber, D., 1975. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. J. Math. Psychol. 12, 387–415.

Barnes, N., 2013. An Overview of Vision Processing in Implantable Prosthetic Vision, in: Image Processing (ICIP), 2013 20th IEEE International Conference on. IEEE, pp. 1532–1535.

Bay, H., Ess, A., Tuytelaars, T., Van Gool, L., 2008. Speeded-up robust features (SURF). Comput. Vis. Image Underst. 110, 346–359.

Beauregard, S., 2006. A helmet-mounted pedestrian dead reckoning system, in: Applied Wearable Computing (IFAWC), 2006 3rd International Forum on. pp. 1–11.

Beauregard, S., Haas, H., 2006. Pedestrian dead reckoning: A basis for personal positioning, in: Proceedings of the 3rd Workshop on Positioning, Navigation and Communication (WPNC'06). pp. 27–35.

Begault, D.R., 1994. 3-D sound for virtual reality and multimedia. AP Professional, Boston.

Belongie, S., Malik, J., Puzicha, J., 2002. Shape matching and object recognition using shape contexts. Pattern Anal. Mach. Intell. IEEE Trans. On 24, 509–522.

Belongie, S., Mori, G., Malik, J., 2006. Matching with shape contexts, in: Statistics and Analysis of Shapes. Springer, pp. 81–105.

Berman, R.A., Wurtz, R.H., 2008. Exploring the Pulvinar Path to Visual Cortex. Prog. Brain Res. 171, 467–473.

Bigham, J.P., Jayant, C., Ji, H., Little, G., Miller, A., Miller, R.C., Miller, R., Tatarowicz, A., White, B., White, S., 2010a. Vizwiz: nearly real-time answers to visual questions, in: Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology. ACM, pp. 333–342.

Bigham, J.P., Jayant, C., Miller, A., White, B., Yeh, T., 2010b. VizWiz:: LocateIt-enabling blind people to locate objects in their environment, in: Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on. IEEE, pp. 65–72.

Binford, T.O., 1971. Visual perception by computer, in: IEEE Conference on Systems and Control. p. 262.

Binford, T.O., 1995. Body-centered representation and perception, in: Hebert, M., Ponce, J., Boult, T., Gross, A. (Eds.), Object Representation in Computer Vision, Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 207–215.

Bissitt, D., Heyes, A.D., 1980. An application of bio-feedback in the rehabilitation of the blind. Appl. Ergon. 11, 31–33.

Blattner, M.M., Sumikawa, D.A., Greenberg, R.M., 1989. Earcons and icons: Their structure and common design principles. Human–Computer Interact. 4, 11–44.

Blauert, J., Allen, J.S., 1997. Spatial hearing: the psychophysics of human sound localization. The MIT Press.

Blenkhorn, P., Evans, D.G., 1997. A system for enabling blind people to identify landmarks: the sound buoy. IEEE Trans. Rehabil. Eng. 5, 276–278.

Blum, A., Mitchell, T., 1998. Combining labeled and unlabeled data with co-training, in: Proceedings of the Eleventh Annual Conference on Computational Learning Theory. ACM, pp. 92–100.

Bolgiano, D., Meeks, E., J., 1967. A laser cane for the blind. IEEE J. Quantum Electron. 3, 268–268.

Bologna, G., Deville, B., Diego Gomez, J., Pun, T., 2011. Toward local and global perception modules for vision substitution. Neurocomputing 74, 1182–1190.

Bologna, G., Deville, B., Pun, T., 2008. Pairing colored socks and following a red serpentine with sounds of musical instruments, in: Proceedings of the 14th International Conference on Auditory Display (ICAD 2008). Paris, France. CD-ROM.

Bologna, G., Deville, B., Pun, T., 2009. On the use of the auditory pathway to represent image scenes in real-time. Neurocomputing, Brain Inspired Cognitive Systems (BICS 2006) / Interplay Between Natural and Artificial Computation (IWINAC 2007) 72, 839–849.

Bologna, G., Deville, B., Pun, T., 2009. Blind navigation along a sinuous path by means of the See ColOr interface, in: Bioinspired Applications in Artificial and Natural Computation. Springer, pp. 235–243.

Bologna, G., Deville, B., Pun, T., Vinckenbosch, M., 2007. Transforming 3D coloured pixels into musical instrument notes for vision substitution applications. J. Image Video Process. 2007, 8–8.

Boreczky, J.S., Wilcox, L.D., 1998. A hidden Markov model framework for video segmentation using audioand image features.

Borovec, J., 2011. Fusion of heterogeneous data for better positioning of visually impaired pedestrians. PAUL SABATIER UNIVERSITY TOULOUSE III, Toulouse, France.

Borovec, J., Gutierrez, O., Brilhault, A., Kammoun, S., Truillet, P., Jouffrais, C., 2014. Fusion of heterogeneous data for better positioning of visually impaired pedestrians. En Cours Soumission.

Boumenir, Y., 2011. Spatial navigation in real and virtual urban environments: performance and multisensory processing of spatial information in sighted, visually impaired, late and congenitally blind individuals. Université Montpellier II - Sciences et Techniques du Languedoc.

Bournot, M.-C., Lelièvre, F., Tallec, A., 2005. La population en situation de handicap visuel en France. Importance, caractéristiques, incapacités fonctionnelles et difficultés sociales. Observatoire régional de la santé des Pays de la Loire.

Bovik, A.C., Clark, M., Geisler, W.S., 1990. Multichannel texture analysis using localized spatial filters. IEEE Trans. Pattern Anal. Mach. Intell. 12, 55–73.

Brabyn, J., Crandall, W., Gerrey, W., 1993. Talking signs: a remote signage, solution for the blind, visually impaired and reading disabled, in: Proceedings of the 15th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 1993. Presented at the Proceedings of the 15th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 1993, pp. 1309–1310.

Bradley, A.P., 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognit. 30, 1145–1159.

Bradley, D.C., Troyk, P.R., Berg, J.A., Bak, M.J., Cogan, S., Erickson, R., Kufta, C.V., Mascaro, M., McCreery, D., Schmidt, E.M., Towle, V.L., Xu, H., 2005. Visuotopic mapping through a multichannel stimulating implant in primate V1. J. Neurophysiol. 93, 1659–70.

Brady, E., Morris, M.R., Zhong, Y., White, S., Bigham, J.P., 2013. Visual challenges in the everyday lives of blind people, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, pp. 2117–2126.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Brelén, M.E., De Potter, P., Gersdorff, M., Cosnard, G., Veraart, C., Delbeke, J., 2006. Intraorbital implantation of a stimulating electrode for an optic nerve visual prosthesis: case report. J. Neurosurg. 104, 593–597.

Brelén, M.E., Duret, F., Gérard, B., Delbeke, J., Veraart, C., 2005. Creating a meaningful visual perception in blind volunteers by optic nerve stimulation. J. Neural Eng. 2, S22.

Brilhault, A., Kammoun, S., Gutierrez, O., Truillet, P., Jouffrais, C., 2011. Fusion of Artificial Vision and GPS to Improve Blind Pedestrian Positioning. Presented at the 4th IFIP International Conference on New Technologies, Mobility and Security (NTMS), Paris, France, pp. 1 -5.

Brindley, G.S., Lewin, W.S., 1968. The sensations produced by electrical stimulation of the visual cortex. J. Physiol. 196, 479–93.

Bruce, I.W., McKennell, A.C., Walker, E.C., 1991. Blind and partially sighted adults in Britain: the RNIB survey. HM Stationery Office.

Brusnighan, D.A., Strauss, M.G., Floyd, J.M., Wheeler, B.C., 1989. Orientation aid implementing the global positioning system, in: Bioengineering Conference, 1989., Proceedings of the 1989 Fifteenth Annual Northeast. Presented at the Bioengineering Conference, 1989., Proceedings of the 1989 Fifteenth Annual Northeast, pp. 33–34.

Buisson, M., Bustico, A., Chatty, S., Colin, F.-R., Jestin, Y., Maury, S., Mertz, C., Truillet, P., 2002. Ivy: un bus logiciel au service du développement de prototypes de systèmes interactifs, in: Proceedings of the 14th French-Speaking Conference on Human-Computer Interaction (Conférence Francophone Sur l'Interaction Homme-Machine). pp. 223–226.

Bullier, J., 2002. Neural basis of vision. Stevens Handb. Exp. Psychol.

Burrough, P.A., 1986. Principles of geographical information systems for land resources assessment. Geocarto Int. 1, 54–54.

Burrough, P.A., McDonnell, R., Burrough, P.A., McDonnell, R., 1998. Principles of geographical information systems. Oxford university press Oxford.

Burschka, D., Hager, G.D., 2003. V-GPS-Image-Based Control for 3D Guidance Systems. pp. 1789–1795.

Buser, P., Imbert, M., 1987. Vision, Neurophysiologie fonctionnelle IV, Hermann. Paris.

Caddeo, P., Fornara, F., Nenci, A.M., Piroddi, A., 2006. Wayfinding tasks in visually impaired people: the role of tactile maps. Cogn. Process. 7, 168–169.

Calkins, D.J., 2001. Seeing with S cones. Prog. Retin. Eye Res. 20, 255–287.

Capelle, C., Trullemans, C., Arno, P., Veraart, C., 1998. A real-time experimental prototype for enhancement of vision rehabilitation using auditory substitution. Biomed. Eng. IEEE Trans. On 45, 1279–1293.

Carabellese, C., Appollonio, I., Rozzini, R., Bianchetti, A., Frisoni, G.B., Frattola, L., Trabucchi, M., 1993. Sensory impairment and quality of life in a community elderly population. J. Am. Geriatr. Soc. 41, 401–407.

Ceranka, S., 2002. Sensor Fusion Algorithms for Pedestrian Location. pp. 1343–1348.

Ceranka, S., Niedzwiecki, M., 2003. Application of particle filtering in navigation system for blind. pp. 495–8.

Cha, K., Horch, K.W., Normann, R.A., 1992. Mobility performance with a pixelized vision system. Vision Res. 32, 1367–1372.

Chalupa, L.M., 1977. A review of cat and monkey studies implicating the pulvinar in visual function. Behav. Biol. 20, 149–167.

Chang, S.F., Ellis, D., Jiang, W., Lee, K., Yanagawa, A., Loui, A.C., Luo, J., 2007. Large-scale multimodal semantic concept detection for consumer video. ACM New York, NY, USA, pp. 255–264.

Chebat, D.-R., 2010. Un oeil sur la langue : aspects neuro-cognitifs du processus de la navigation chez l'aveugle-né.

Chebat, D.-R., Rainville, C., Kupers, R., Ptito, M., 2007. Tactile-'visual' acuity of the tongue in early blind individuals. Neuroreport 18, 1901–1904.

Cheng, C., O'Leary, B., Stearns, L., Caperna, S., Cho, J., Fan, V., Luthra, A., Sun, A., Tessler, R., Wong, P., 2008. Developing a Real-Time Identify-and-Locate System for the Blind, in: Workshop on Computer Vision Applications for the Visually Impaired.

Chen, X., Yuille, A., 2011. AdaBoost learning for detecting and reading text in city scenes.

Cherkassky, B.V., Goldberg, A.V., Radzik, T., 1996. Shortest paths algorithms: theory and experimental evaluation. Math. Program. 73, 129–174.

Chia, E.M., Wang, J.J., Rochtchina, E., Smith, W., Cumming, R.R., Mitchell, P., 2004. Impact of bilateral visual impairment on health-related quality of life: the Blue Mountains Eye Study. Invest. Ophthalmol. Vis. Sci. 45, 71.

Chincha, R., Tian, Y., 2011. Finding objects for blind people based on SURF features, in: Bioinformatics and Biomedicine Workshops (BIBMW), 2011 IEEE International Conference on. IEEE, pp. 526–527.

Chow, A.Y., Chow, V.Y., Packo, K.H., Pollack, J.S., Peyman, G.A., Schuchard, R., 2004. The artificial silicon retina microchip for the treatment of visionloss from retinitis pigmentosa. Arch. Ophthalmol. 122, 460–469.

Christ, S.L., Lee, D.J., Lam, B.L., Zheng, D.D., Arheart, K.L., 2008. Assessment of the Effect of Visual Impairment on Mortality through Multiple Health Pathways: Structural Equation Modeling. Invest Ophthalmol Vis Sci 49, 3318–3323.

Chumkamon, S., Tuvaphanthaphiphat, P., Keeratiwintakorn, P., 2008. A blind navigation system using RFID for indoor environments, in: Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, 2008. ECTI-CON 2008. 5th International Conference on. IEEE, pp. 765–768.

Chum, O., Philbin, J., Isard, M., Zisserman, A., 2007. Scalable near identical image and shot detection, in: Proceedings of the 6th ACM International Conference on Image and Video Retrieval. ACM, pp. 549–556.

Chum, O., Zisserman, A., 2007. An exemplar model for learning object classes, in: Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on. IEEE, pp. 1–8.

Ciresan, D., Meier, U., Schmidhuber, J., 2012. Multi-column deep neural networks for image classification. pp. 3642–3649.

Clark-Carter, D.D., Heyes, A.D., Howarth, C.I., 1986. The efficiency and walking speed of visually impaired people. Ergonomics 29, 779–789.

Clarke, K.C., McLafferty, S.L., Tempalski, B.J., 1996. On epidemiology and geographic information systems: a review and discussion of future directions. Emerg. Infect. Dis. 2, 85.

Cohen, E.D., 2007. Prosthetic interfaces with the visual system: biological issues. J. Neural Eng. 4, R14.

Congdon NG, Friedman DS, Lietman T, 2003. Important causes of visual impairment in the world today. JAMA 290, 2057–2060.

Coughlan, J., Manduchi, R., 2007. Functional assessment of a camera phone-based wayfinding system operated by blind users, in: Conference of IEEE Computer Society and the Biological and Artificial Intelligence Society (IEEE-BAIS), Research on Assistive Technologies Symposium (RAT'07).

Coughlan, J., Manduchi, R., Shen, H., 2006. Cell phone-based wayfinding for the visually impaired. Proc IMV 2006.

Coultrip, R., Granger, R., Lynch, G., 1992. A cortical model of winner-take-all competition via lateral inhibition. Neural Netw. 5, 47–54.

Cox, D., Pinto, N., 2011. Beyond simple features: A large-scale feature search approach to unconstrained face recognition. pp. 8–15.

Craig, J.C., 1981. Tactile letter recognition: Pattern duration and modes of pattern generation. Percept. Psychophys. 30, 540–546.

Crandall, W., Brabyn, J., Bentzen, B.L., Myers, L., 1999. Remote infrared signage evaluation for transit stations and intersections. J. Rehabil. Res. Dev. 36.

Cronly-Dillon, J., Persaud, K.C., Blore, R., 2000. Blind subjects construct conscious mental images of visual scenes encoded in musical form. Proc. R. Soc. B Biol. Sci. 267, 2231–2238.

Cronly-Dillon, J., Persaud, K., Gregory, R.P.F., 1999. The perception of visual images encoded in musical form: a study in cross-modality information transfer. Proc. R. Soc. Lond. B Biol. Sci. 266, 2427–2433.

Cross, G.R., Jain, A.K., 1981. Markov random field texture models. pp. 597–602.

Cui, Y., Schuon, S., Chan, D., Thrun, S., Theobalt, C., 2010. 3D shape scanning with a time-of-flight camera, in: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. pp. 1173–1180.

Curcio, C.A., Sloan, K.R., Kalina, R.E., Hendrickson, A.E., 1990. Human photoreceptor topography. J. Comp. Neurol. 292, 497–523.

Dacey, D.M., 1996. Circuitry for color coding in the primate retina. Proc. Natl. Acad. Sci. 93, 582–588.

Dacey, D.M., Lee, B.B., 1994. The'blue-on'opponent pathway in primate retina originates from a distinct bistratified ganglion cell type. Nature 367, 731–735.

Dacey, D., Packer, O.S., Diller, L., Brainard, D., Peterson, B., Lee, B., 2000. Center surround receptive field structure of cone bipolar cells in primate retina. Vision Res. 40, 1801–1811.

Dagnelie, G., 2008. Psychophysical Evaluation for Visual Prosthesis. Annu. Rev. Biomed. Eng. 10, 339–368.

Dagnelie, G., 2011. Visual prosthetics: physiology, bioengineering, rehabilitation. Springer.

Dakopoulos, D., Bourbakis, N.G., 2010. Wearable Obstacle Avoidance Electronic Travel Aids for Blind: A Survey. IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. 40, 25–35.

Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection, in: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. pp. 886–893.

Damaschini, R., Legras, R., Leroux, R., Farcy, R., 2005. Electronic Travel Aid for blind people. Assist. Technol. Virtuality Real. 16, 251–255.

Dan, Y., Atick, J.J., Reid, R.C., 1996. Efficient Coding of Natural Scenes in the Lateral Geniculate Nucleus: Experimental Test of a Computational Theory. J. Neurosci. 16, 3351–3362.

Daugman, J.G., 1985. Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. Opt. Soc. Am. J. Opt. Image Sci. 2, 1160–1169.

Davis, J., Goadrich, M., 2006. The relationship between Precision-Recall and ROC curves. pp. 233–240.

Delbeke, J., Oozeer, M., Veraart, C., 2003. Position, size and luminosity of phosphenes generated by direct optic nerve stimulation. Vision Res. 43, 1091–1102.

Delbeke, J., Wanet-Defalque, M. c., Gérard, B., Troosters, M., Michaux, G., Veraart, C., 2002. The Microsystems Based Visual Prosthesis for Optic Nerve Stimulation. Artif. Organs 26, 232–234.

Del Bimbo, A., Pala, P., Santini, S., 1996. Image retrieval by elastic matching of shapes and image patterns, in: Multimedia Computing and Systems, 1996., Proceedings of the Third IEEE International Conference on. pp. 215–218.

DeLong, E.R., DeLong, D.M., Clarke-Pearson, D.L., 1988. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 44, 837–845.

Delorme, A., Gautrais, J., van Rullen, R., Thorpe, S., 1999. SpikeNET: A simulator for modeling large networks of integrate and fire neurons. Neurocomputing 26, 989–996.

Delorme, A., Perrinet, L., Thorpe, S.J., 2001. Networks of integrate-and-fire neurons using Rank Order Coding B: Spike timing dependent plasticity and emergence of orientation selectivity. Neurocomputing, Computational Neuroscience: Trends in Research 2001 38–40, 539–545.

Delorme, A., Thorpe, S.J., 2003. SpikeNET: an event-driven simulation package for modelling large networks of spiking neurons. Netw. Comput. Neural Syst. 14, 613–627.

Deng, Y., Manjunath, B.S., Kenney, C., Moore, M.S., Shin, H., 2001. An efficient color representation for image retrieval. IEEE Trans. Image Process. 10, 140–147.

Denham, J., Leventhal, J., McComas, H., 2004. Getting from Point A to Point B: A review of two GPS systems.

Denis, G., Jouffrais, C., Vergnieux, V., Macé, M., 2013. Human faces detection and localization with simulated prosthetic vision, in: CHI'13 Extended Abstracts on Human Factors in Computing Systems. pp. 61–66.

Denis, G., Macé, M.J.-M., Jouffrais, C., 2012. Simulated prosthetic vision: object recognition and localization approach, in: Proceedings of the 4th International Conference on Neuroprosthetic Devices (ICNPD 2012). Presented at the 4th International Conference on Neuroprosthetic Devices (ICNPD 2012).

Deville, B., Bologna, G., Vinckenbosch, M., Pun, T., 2008. Guiding the focus of attention of blind people with visual saliency, in: Workshop on Computer Vision Applications for the Visually Impaired.

Deville, B., Bologna, G., Vinckenbosch, M., Pun, T., 2009. See color: Seeing colours with an orchestra, in: Human Machine Interaction. Springer, pp. 251–279.

Dhond, U.R., Aggarwal, J.K., 1989. Structure from stereo-a review. Syst. Man Cybern. IEEE Trans. On 19, 1489–1510.

DiCiccio, T.J., Efron, B., 1996. Bootstrap confidence intervals. Stat. Sci. 189–212.

Dijkstra, E.W., 1959. A note on two problems in connexion with graphs. Numer. Math. 1, 269–271.

Dobelle, W.H., 2000. Artificial vision for the blind by connecting a television camera to the visual cortex. ASAIO J. 46, 3–9.

Dobelle, W.H., Mladejovsky, M.G., Girvin, J.P., 1974. Artificial vision for the blind: electrical stimulation of visual cortex offers hope for a functional prosthesis. Science 183, 440–444.

Dobelle, W.H., Quest, D.O., Antunes, J.L., Roberts, T.S., Girvin, J.P., 1979. Artificial vision for the blind by electrical stimulation of the visual cortex. Neurosurgery 5, 521–527.

Dorfman, D.D., Alf, E., 1969. Maximum-likelihood estimation of parameters of signal-detection theory and determination of confidence intervals—Rating-method data. J. Math. Psychol. 6, 487–496.

Dowling, J.A., Maeder, A., Boles, W., 2004. Mobility enhancement and assessment for a visual prosthesis, in: Medical Imaging 2004. pp. 780–791.

Dramas, F., 2010. Localisation d'objets pour les non-voyants : augmentation sensorielle et neuroprothèse (phd). Université de Toulouse, Université Toulouse III - Paul Sabatier.

Dramas, F., Thorpe, S.J., Jouffrais, C., 2010. Artificial Vision For The Blind: A Bio-Inspired Algorithm For Objects And Obstacles Detection. Int. J. Image Graph. Vol 10 No 4 531–544.

Draper, B.A., Bins, J., Baek, K., 1999. ADORE: adaptive object recognition, in: Computer Vision Systems. Springer, pp. 522–537.

Dunai, L., Fajarnes, G.P., Praderas, V.S., Garcia, B.D., Lengua, I.L., 2010. Real-time assistance prototype — A new navigation aid for blind people, in: IECON 2010-36th Annual Conference on IEEE Industrial Electronics Society. pp. 1173–1178.

Duret, F., Brelén, M.E., Lambert, V., Gérard, B., Delbeke, J., Veraart, C., 2006. Object localization, discrimination, and grasping with the optic nerve visual prosthesis. Restor. Neurol. Neurosci. 24, 31–40.

Durette, B., 2009. Traitement du signal pour les prothèses visuelles: approche biométrique et sensori-motrice. Université Joseph-Fourier - Grenoble I.

Durette, B., Louveton, N., Alleysson, D., Hérault, J., 2008. Visuo-auditory sensory substitution for mobility assistance: testing TheVIBE. Presented at the Computer Vision Applications for the Visually Impaired (CVAVI 2008).

Durrant-Whyte, H., Bailey, T., 2006. Simultaneous localization and mapping: part I. Robot. Autom. Mag. IEEE 13, 99–110.

Eakins, J., Graham, M., Newcastle, U. of N. at, 1999. Content-based image retrieval. University of Northumbria at Newcastle.

Eakins, J.P., 1996. Automatic image content retrieval-are we getting anywhere?, in: ELVIRA-PROCEEDINGS-. pp. 121–134.

Eakins, J.P., 1998. Techniques for image retrieval. Libr. Inf. Brief. 1–15.

Eckmiller, R., 1997. Learning Retina Implants with Epiretinal Contacts. Ophthalmic Res. 29, 281–289.

Eickeler, S., Müller, S., 1999. Content-based video indexing of TV broadcast news using hiddenMarkov models.

Epstein, W., 1985. Amodal information and transmodal perception, in: Electronic Spatial Sensing for the Blind. Springer, pp. 421–430.

Everingham, M., Gool, L., Williams, C.K.I., Winn, J., Zisserman, A., 2009. The Pascal Visual Object Classes (VOC) Challenge. Int. J. Comput. Vis. 88, 303–338.

Fabiani, P., 1996. Représentation dynamique de l'incertain et stratégie de perception pour un système autonome en environnement évolutif.

Faraggi, D., Reiser, B., 2002. Estimation of the area under the ROC curve. Stat. Med. 21, 3093–3106.

Farcy, R., Leroux, R., Damaschini, R., Legras, R., Bellik, Y., Jacquet, C., Pardo, P., 2003. Laser telemetry to improve the mobility of blind people: Report of the 6 month training course, in: Proceedings of the 1st International Conference on Smart Homes and Health Telematics. pp. 113–115.

Farcy, R., Leroux, R., Jucha, A., Damaschini, R., Grégoire, C., Zogaghi, A., 2006. Electronic travel aids and electronic orientation aids for blind people: technical, rehabilitation and everyday life points of view, in: Conference & Workshop on Assistive Technologies for People with Vision & Hearing Impairments Technology for Inclusion. p. 12.

Farcy, R., Leroux, R., Jucha, A., Damaschini, R., Grégoire, C., Zogaghi, A., 2006. Electronic travel aids and electronic orientation aids for blind people: Technical, rehabilitation and everyday life points of view. CVHI.

Fatourechi, M., Ward, R.K., Mason, S.G., Huggins, J., Schlögl, A., Birch, G.E., 2008. Comparison of Evaluation Metrics in Classification Applications with Imbalanced Datasets. IEEE, pp. 777–782.

Fawcett, T., 2004. ROC graphs: Notes and practical considerations for researchers. Mach. Learn. 31, 1–38.

Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recognit. Lett. 27, 861–874.

Feliz Alonso, R., Zalama Casanova, E., Gómez García-Bermejo, J., 2009. Pedestrian tracking using inertial sensors.

Felson, D.T., Anderson, J.J., Hannan, M.T., Milton, R.C., 1989. Impaired vision and hip fracture: the Framingham Study. J. Am. Geriatr. Soc.

Fernández, E., Pelayo, F., Romero, S.F., Bongard, M., Marin, C., Alfaro, A., Merabet, L.B., 2005. Development of a cortical visual neuroprosthesis for the blind: the relevance of neuroplasticity. J. Neural Eng. 2, R1–12.

Field, D.J., Olmos, A., 2007. Does spatial invariance result from insensitivity to change? J. Vis. 7.

Fischer, C., Muthukrishnan, K., Hazas, M., Gellersen, H., 2008. Ultrasound-aided pedestrian dead reckoning for indoor navigation, in: Proceedings of the First ACM International Workshop on Mobile Entity Localization and Tracking in GPS-Less Environments. pp. 31–36.

Flach, P.A., 2003. The geometry of ROC space: understanding machine learning metrics through ROC isometrics, in: ICML. pp. 194–201.

Foran, S., Wang, J.J., Rochtchina, E., Mitchell, P., 2000. Projected number of Australians with visual impairment in 2000 and 2030. Clin. Experiment. Ophthalmol. 28, 143–145.

Förstner, W., 1994. A framework for low level feature extraction. Springer-Verlag.

Foster, A., Resnikoff, S., 2005. The impact of Vision 2020 on global blindness. Eye 19, 1133–1135.

Foulke, E., 1982. Perception, cognition and the mobility of blind pedestrians. Spat. Abil. Dev. Physiol. Found. 55–76.

Freeman, W.T., Adelson, E.H., 1991. The design and use of steerable filters. Pattern Anal. Mach. Intell. IEEE Trans. On 13, 891–906.

Frey, A., Daquet, A., Poitrenaud, S., Tijus, C., Fremiot, M., Formosa, M., Prod homme, L., Mandelbrojt, J., Timsit-Berthier, M., Bootz, P., Hautbois, X., Besson, M., 2009. Pertinence cognitive des unités sémiotiques temporelles.

Frick, K.D., Foster, A., 2003. The magnitude and cost of global blindness: an increasing problem that can be alleviated. Am. J. Ophthalmol. 135, 471–476.

Friedman, D.S., Freeman, E., Munoz, B., Jampel, H.D., West, S.K., 2007. Glaucoma and Mobility Performance: The Salisbury Eye Evaluation Project. Ophthalmology 114, 2232–2237.e1.

Gaunet, F., Briffault, X., 2005. Exploring the functional specifications of a localized wayfinding verbal aid for blind pedestrians: Simple and structured urban areas. Hum.-Comput. Interact. 20, 267–314.

Gemert, J.C. van, Snoek, C.G.M., Veenman, C.J., Smeulders, A.W.M., 2006. The influence of cross-validation on video classification performance. ACM, Santa Barbara, CA, USA, pp. 695–698.

Gerrits, M., Bekaert, P., 2006. Local stereo matching with segmentation-based outlier rejection, in: Computer and Robot Vision, 2006. The 3rd Canadian Conference on. pp. 66–66.

Geruschat, D.R., Turano, K.A., 2007. Estimating the Amount of Mental Effort Required for Independent Mobility: Persons with Glaucoma. Invest Ophthalmol Vis Sci 48, 3988–3994.

Gibson, J.J., 1966. The senses considered as perceptual systems.

Gibson, J.J., 1986. The ecological approach to visual perception. Routledge.

Goldish, L.H., Taylor, H.E., 1974. The Optacon: A Valuable Device for Blind Persons. New Outlook Blind 68, 49–56.

Golledge, J.M.L., 1991. Designing a personal guidance system to aid navigation without sight: progress on the GIS component. Int. J. Geogr. Inf. Sci. 5, 373–395.

Golledge, R.G., 1993. Geography and the disabled: a survey with special reference to vision impaired and blind populations. Trans. Inst. Br. Geogr. 63–85.

Golledge, R.G., Dougherty, V., Bell, S., 1995. Acquiring spatial knowledge: Survey versus route-based knowledge in unfamiliar environments. Ann. Assoc. Am. Geogr. 85, 134–158.

Golledge, R.G., Klatzky, R.L., Loomis, J.M., Speigle, J., Tietz, J., 1998. A geographical information system for a GPS based personal guidance system. Int. J. Geogr. Inf. Sci. 12, 727–749.

Golledge, R., Klatzky, R., Loomis, J., Marston, J., 2004. Stated preferences for components of a personal guidance system for nonvisual navigation. J. Vis. Impair. Blind. JVIB 98.

Gomez Valencia, J.D., 2014. A computer-vision based sensory substitution device for the visually impaired (See ColOr). University of Geneva.

Gonzalez-Mora, J.L., Rodriguez-Hernandez, A., Burunat, E., Martin, F., Castellano, M.A., 2006. Seeing the world by hearing: Virtual Acoustic Space (VAS) a new space perception system for blind people., in: Information and Communication Technologies, 2006. ICTTA '06. 2nd. Presented at the Information and Communication Technologies, 2006. ICTTA '06. 2nd, pp. 837–842.

Gonzalez-Mora, J.L., Rodriguez-Hernandez, A., Rodriguez-Ramos, L.F., Díaz-Saco, L., Sosa, N., 1999. Development of a new space perception system for blind people, based on the creation of a virtual acoustic space, in: Engineering Applications of Bio-Inspired Artificial Neural Networks. Springer, pp. 321–330.

Goodchild, M.F., 1991. Geographic information systems. J. Retail. 67, 3–15.

Gotlieb, C.C., Kreyszig, H.E., 1990. Texture descriptors based on co-occurrence matrices. Comput. Vis. Graph. Image Process. 51, 70–86.

Green, D.M., Swets, J.A., 1966. Signal detection theory and psychophysics. Wiley New York.

Grieve, K.L., Acuña, C., Cudeiro, J., 2000. The primate pulvinar nuclei: vision and action. Trends Neurosci. 23, 35–39.

Grumet, A.E., Wyatt Jr., J.L., Rizzo III, J.F., 2000. Multi-electrode stimulation and recording in the isolated retina. J. Neurosci. Methods 101, 31–42.

Guarniero, G., 1974. Experience of tactile vision. Perception 3, 101–104.

Gude, R., Østerby, M., Soltveit, S., n.d. Blind Navigation and Object Recognition.

Gudivada, V.N., Raghavan, V.V., 1995. Content based image retrieval systems. Computer 28, 18–22.

Guillery, R.W., Sherman, S.M., 2002. Thalamic Relay Functions and Their Role in Corticocortical Communication: Generalizations from the Visual System. Neuron 33, 163–175.

Guth, D.A., Rieser, J.J., 1997. Perception and the control of locomotion by blind and visually impaired pedestrians. Found. Orientat. Mobil. 2, 9–38.

Hahne, U., Alexa, M., 2008. Combining time-of-flight depth and stereo images without accurate extrinsic calibration. Int. J. Intell. Syst. Technol. Appl. 5, 325–333.

Hallum, L.E., Suaning, G.J., Taubman, D.S., Lovell, N.H., 2005. Simulated prosthetic visual fixation, saccade, and smooth pursuit. Vision Res. 45, 775–788.

Hamilton, R.H., Pascual-Leone, A., 1998. Cortical plasticity associated with Braille learning. Trends Cogn. Sci. 2, 168–174.

Hanley, J.A., Hajian-Tilaki, K.O., 1997. Sampling variability of nonparametric estimates of the areas under receiver operating characteristic curves: an update. Acad. Radiol. 4, 49–58.

Hanley, J.A., McNeil, B.J., 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143, 29–36.

Hanneton, S., Auvray, M., Durette, B., 2010. The Vibe: a versatile vision-to-audition sensory substitution device. Appl. Bionics Biomech. 7, 269–276.

Haque, S., Kulik, L., Klippel, A., 2007. Algorithms for reliable navigation and wayfinding, in: Spatial Cognition V Reasoning, Action, Interaction. Springer, pp. 308–326.

Haralick, R.M., Shanmugam, K., Dinstein, I.H., 1973. Textural features for image classification. IEEE Trans. Syst. Man Cybern. 3, 610–621.

Harris, D., Whitney, G., 1995. Smart signs, in: 7th International Conference on Mobility and Transport for Elderly and Disabled People.

Hasanuzzaman, F.M., Yang, X., Tian, Y., 2011. Robust and effective component-based banknote recognition by SURF features, in: Wireless and Optical Communications Conference (WOCC), 2011 20th Annual. IEEE, pp. 1–6.

Havik, E.M., Steyvers, F.J., van der Velde, H., Pinkster, J.C., Kooijman, A.C., 2010. Design and Evaluation of a Protocol to Assess Electronic Travel Aids for Persons Who Are Visually Impaired. J. Vis. Impair. Blind. 104.

Heath, M.D., Sarkar, S., Sanocki, T., Bowyer, K.W., 1997. A robust visual method for assessing the relative performance of edge-detection algorithms. Pattern Anal. Mach. Intell. IEEE Trans. On 19, 1338–1359.

Helal, A., Moore, S.E., Ramachandran, B., 2001. Drishti: An integrated navigation system for visually impaired and disabled, in: Wearable Computers, 2001. Proceedings. Fifth International Symposium on. IEEE, pp. 149–156.

Hendry, S.H., Reid, R.C., 2000. The koniocellular pathway in primate vision. Annu. Rev. Neurosci. 23, 127–153.

Henze, N., Heuten, W., Boll, S., 2006. Non-intrusive somatosensory navigation support for blind pedestrians. pp. 459–464.

He, X., Kim, J., Barnes, N., 2012. An face-based visual fixation system for prosthetic vision, in: Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE. IEEE, pp. 2981–2984.

Hirota, G., Chen, D.T., Garrett, W.F., Livingston, M.A., 1996. Superior augmented reality registration by integrating landmark tracking and magnetic tracking, in: Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques. ACM, pp. 429–438.

Hirsh, I.J., 1988. Auditory perception and speech. Handb. Exp. Psychol. 377–408.

Horne, L., Barnes, N., McCarthy, C., He, X., 2012. Image segmentation for enhancing symbol recognition in prosthetic vision, in: Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE. IEEE, pp. 2792–2795.

Hosni, A., Bleyer, M., Gelautz, M., Rhemann, C., 2009. Local stereo matching using geodesic support weights, in: 2009 16th IEEE International Conference on Image Processing (ICIP). Presented at the 2009 16th IEEE International Conference on Image Processing (ICIP), pp. 2093–2096.

Hoyle, B.S., Dodds, S., 2006. The UltraCane mobility aid at work. From training programs to case studies, in: Conference and Workshop on Assistive Technologies for People with Vision & Hearing Impairments, Technology for Inclusion, CVHI.

Huang, J., Kumar, S.R., Mitra, M., Zhu, W.J., 2001. Image indexing using color correlograms. Google Patents.

Huang, J., Ling, C.X., 2005. Using AUC and accuracy in evaluating learning algorithms. IEEE Trans. Knowl. Data Eng. 17, 299–310.

Hub, A., Diepstraten, J., Ertl, T., 2005. Augmented Indoor Modeling for Navigation Support for the Blind., in: CPSN. pp. 54–62.

Hub, A., Hartter, T., Ertl, T., 2006a. Interactive localization and recognition of objects for the blind, in: California State University, Northridge Center on Disabilities' 21st Annual International Technology and Persons with Disabilities Conference.

Hub, A., Hartter, T., Ertl, T., 2006b. Interactive tracking of movable objects for the blind on the basis of environment models and perception-oriented object recognition

methods, in: Proceedings of the 8th International ACM SIGACCESS Conference on Computers and Accessibility. pp. 111–118.

Hubel, D., 1994. L'oeil, le cerveau et la vision. Paris Pour Sci.

Hughes, B., 2001. Active artificial echolocation and the nonvisual perception of aperture passability. Hum. Mov. Sci. 20, 371–400.

Humayun, M.S., de Juan Jr., E., Weiland, J.D., Dagnelie, G., Katona, S., Greenberg, R., Suzuki, S., 1999. Pattern electrical stimulation of the human retina. Vision Res. 39, 2569–2576.

Humayun, M.S., Weiland, J.D., Fujii, G.Y., Greenberg, R., Williamson, R., Little, J., Mech, B., Cimmarusti, V., Van Boemel, G., Dagnelie, G., de Juan Jr., E., 2003. Visual perception in a blind subject with a chronic microelectronic retinal prosthesis. Vision Res. 43, 2573–2581.

Hussmann, S., Ringbeck, T., Hagebeuker, B., 2008. A performance review of 3D TOF vision systems in comparison to stereo vision systems. Stereo Vis. 103–120.

IAPB, 2010. International Agency for the Prevention of Blindness - 2010 Report.

Ip, S.P.S., Leung, Y.F., Mak, W.P., 2000. Depression in institutionalised older people with impaired vision. Int. J. Geriatr. Psychiatry 15, 1120–1124.

Ivanov, R., 2011. Algorithm for blind navigation along a GPS track., in: CompSysTech. pp. 372–379.

Ivers, R.Q., Norton, R., Cumming, R.G., Butler, M., Campbell, A.J., 2000. Visual Impairment and Hip Fracture. Am. J. Epidemiol. 152, 633–639.

Jacobs, J.M., Hammerman-Rozenberg, R., Maaravi, Y., Cohen, A., Stessman, J., 2005. The impact of visual impairment on health, function and mortality. Aging Clin. Exp. Res. 17, 281–286.

Jacobson, R.D., Kitchin, R.M., 1997. GIS and people with visual impairments or blindness: exploring the potential for education, orientation, and navigation. Trans. GIS 2, 315–332.

Jacquot, A., Sturm, P., Ruch, O., 2005. Adaptative Tracking of Non-Rigid Objects Based on Color Histograms and Automatic Parameter Selection. IEEE, pp. 103–109.

Jafri, R., Ali, S.A., Arabnia, H.R., Fatima, S., 2013. Computer vision-based object recognition for the visually impaired in an indoors environment: a survey. Vis. Comput.

Jaimes, A., Sebe, N., 2007. Multimodal human–computer interaction: A survey. Comput. Vis. Image Underst. 108, 116–134.

Jain, A.K., Duin, R.P.W., Mao, J., 2000. Statistical pattern recognition: A review. Pattern Anal. Mach. Intell. IEEE Trans. On 22, 4–37.

Jansson, G., 1983. Tactile guidance of movement. Int. J. Neurosci. 19, 37–46.

Javaheri, M., Hahn, D.S., Lakhanpal, R.R., Weiland, J.D., Humayun, M.S., 2006. Retinal prostheses for the blind. Ann. Acad. Med. Singapore 35, 137–144.

Javed, O., Ali, S., Shah, M., 2005. Online detection and classification of moving objects using progressively improving detectors, in: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. IEEE, pp. 696–701.

Jianguo Zhang, Marszalek, M., Lazebnik, S., Schmid, C., 2006. Local Features and Kernels for Classification of Texture and Object Categories: A Comprehensive Study. Presented at the Computer Vision and Pattern Recognition Workshop, 2006. CVPRW '06. Conference on, p. 13.

Jimenez, A.R., Seco, F., Prieto, C., Guevara, J., 2009. A comparison of pedestrian dead-reckoning algorithms using a low-cost MEMS IMU, in: Intelligent Signal Processing, 2009. WISP 2009. IEEE International Symposium on. pp. 37–42.

Joachims, T., 1999. SVMLight Support Vector Machine. SVM-Light Support Vector Mach. Httpsvmlight Joachims Org Univ. Dortm.

Johnson, L.A., Higgins, C.M., 2006. A navigation aid for the blind using tactile-visual sensory substitution, in: Engineering in Medicine and Biology Society, 2006. EMBS'06. 28th Annual International Conference of the IEEE. IEEE, pp. 6289–6292.

Johnson, M.H., 2005. Subcortical face processing. Nat. Rev. Neurosci. 6, 766–774.

Jones, J.P., Palmer, L.A., 1987. An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. J. Neurophysiol. 58, 1233–1258.

Kaczmarek, K.A., Haase, S.J., 2003. Pattern identification as a function of stimulation on a fingertip-scanned electrotactile display. Neural Syst. Rehabil. Eng. IEEE Trans. On 11, 269–275.

Kaczmarek, K.A., Tyler, M.E., Bach-y-Rita, P., 1997. Pattern identification on a fingertip-scanned electrotactile display, in: Engineering in Medicine and Biology Society, 1997. Proceedings of the 19th Annual International Conference of the IEEE. pp. 1694–1696.

Kaczmarek, K.A., Webster, J.G., Bach-y-Rita, P., Tompkins, W.J., 1991. Electrotactile and vibrotactile displays for sensory substitution systems. Biomed. Eng. IEEE Trans. On 38, 1–16.

Kalal, Z., Matas, J., Mikolajczyk, K., 2009. Online learning of robust object detectors during unstable tracking, in: Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on. IEEE, pp. 1417–1424.

Kammoun, S., 2009. Conception et évaluation de l'interface en mobilité d'un système de guidage par GNSS et vision artificielle pour non-voyants (Rapport de Master 2 recherche Informatique et Télécommunication). l'Université Paul Sabatier (UPS) Toulouse III, Toulouse, France.

Kammoun, S., 2013. Assistance à la navigation pour les non-voyants : Vers un positionnement, un SIG et un suivi adaptés. Université de Toulouse - Paul Sabatier, Toulouse, France.

Kammoun, S., Dramas, F., Oriola, B., Jouffrais, C., 2010. Route Selection Algorithm for Blind Pedestrian. Presented at the International Conference on Control, Automation and Systems, IEEE, KINTEX, Gyeonggi-do, Korea, pp. 2223– 2228.

Kammoun, S., Macé, M.J.-M., Oriola, B., Jouffrais, C., 2012. Towards a geographic information system facilitating navigation of visually impaired users, in: Computers Helping People with Special Needs. Springer, pp. 521–528.

Kaplan, E., Hegarty, C., 2006. Understanding GPS: Principles and Applications Second Edition. Artech House.

Kaplan, L.M., Murenzi, R., Namuduri, K.R., 1997. Fast texture database retrieval using extended fractal features. pp. 162–173.

Kato, T., 1992. Database architecture for content-based image retrieval. pp. 112–123.

Katz, B.F.G., Dramas, F., Parseihian, G., Gutierrez, O., Kammoun, S., Brilhault, A., Brunet, L., Gallay, M., Oriola, B., Auvray, M., Truillet, P., Denis, M., Thorpe, S., Jouffrais, C., 2012a. NAVIG: Guidance system for the visually impaired using virtual augmented reality. Technol. Disabil.

Katz, B.F.G., Kammoun, S., Parseihian, G., Gutierrez, O., Brilhault, A., Auvray, M., Truillet, P., Denis, M., Thorpe, S., Jouffrais, C., 2012b. NAVIG: Augmented reality guidance system for the visually impaired. Virtual Real.

Katz, B.F.G., Truillet, P., Thorpe, S., Jouffrais, C., 2010. NAVIG: Navigation Assisted by Artificial Vision and GNSS. Presented at the Workshop on Multimodal Location Based Techniques for Extreme Navigation, Pervasive, in press, Helsinki, Finland.

Katz, B., Rio, E., Picinali, L., 2010. LIMSI Spatialization Engine. Inter Deposit Digital Number: IDDN.FR.001.340014.000.S.P.2010.000.31235.

Kawai, Y., Tomita, F., 2002. A support system for visually impaired persons to understand three-dimensional visual information using acoustic interface, in: Pattern Recognition, 2002. Proceedings. 16th International Conference on. IEEE, pp. 974–977.

Kawamura, S., Tachibanaki, S., 2008. Rod and cone photoreceptors: molecular basis of the difference in their physiology. Comp. Biochem. Physiol. A. Mol. Integr. Physiol. 150, 369–377.

Kay, L., 1974. A sonar aid to enhance spatial perception of the blind: engineering design and evaluation. Radio Electron. Eng. 44, 605–627.

Kay, L., 1984. Electronic aids for blind persons: an interdisciplinary subject. Phys. Sci. Meas. Instrum. Manag. Educ. - Rev. IEE Proc. A 131, 559–576.

Kim, T.-K., Kittler, J., Cipolla, R., 2007. Discriminative learning and recognition of image set classes using canonical correlations. Pattern Anal. Mach. Intell. IEEE Trans. On 29, 1005–1018.

Klaver CW, Wolfs RW, Vingerling JR, Hofman A, de Jong PM, 1998. Age-specific prevalence and causes of blindness and visual impairment in an older population: The rotterdam study. Arch. Ophthalmol. 116, 653–658.

Klein, B.E.K., Klein, R., Lee, K.E., Cruickshanks, K.J., 1998. Performance-based and self-assessed measures of visual function as related to history of falls, hip fractures, and measured gait time: The beaver dam eye study. Ophthalmology 105, 160–164.

Knierim, J.J., Van Essen, D.C., 1992. Neuronal responses to static texture patterns in area V1 of the alert macaque monkey. J. Neurophysiol. 67, 961–980.

Krishna, S., Colbry, D., Black, J., Balasubramanian, V., Panchanathan, S., 2008. A systematic requirements analysis and development of an assistive device to enhance the social interaction of people who are blind or visually impaired, in: Workshop on Computer Vision Applications for the Visually Impaired.

Krishna, S., Little, G., Black, J., Panchanathan, S., 2005. A wearable face recognition system for individuals with visual impairments, in: Proceedings of the 7th International ACM SIGACCESS Conference on Computers and Accessibility. pp. 106–113.

Kuffler, S.W., 1953. Discharge patterns and functional organization of mammalian retina. J Neurophysiol 16, 37–68.

Kulyukin, V., Kutiyanawala, A., 2010. From ShopTalk to ShopMobile: vision-based barcode scanning with mobile phones for independent blind grocery shopping, in: Proceedings of the 2010 Rehabilitation Engineering and Assistive Technology Society of North America Conference (RESNA 2010), Las Vegas, NV.

Kupers, R., Fumal, A., de Noordhout, A.M., Gjedde, A., Schoenen, J., Ptito, M., 2006. Transcranial magnetic stimulation of the visual cortex induces somatotopically organized qualia in blind subjects. Proc. Natl. Acad. Sci. 103, 13256–13260.

Kupers, R., Pietrini, P., Ricciardi, E., Ptito, M., 2011. The Nature of Consciousness in the Visually Deprived Brain. Front. Psychol. 2.

Kutiyanawala, A., Kulyukin, V., 2010. Eyes-free barcode localization and decoding for visually impaired mobile phone users, in: 2010 International Conference on Image Processing, Computer Vision and Pattern Recognition. pp. 130–135.

Lampert, C.H., Blaschko, M.B., Hofmann, T., 2008. Beyond sliding windows: Object localization by efficient subwindow search, in: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. IEEE, pp. 1–8.

LaPierre, C.M., 1998. Personal Navigation Systern for the Visually Impaired. Carleton University.

Latecki, L.J., Lakämper, R., 1999. Convexity rule for shape decomposition based on discrete contour evolution. Comput. Vis. Image Underst. 73, 441–454.

Lavrac, N., Flach, P., Zupan, B., 1999. Rule evaluation measures: A unifying view. Springer.

LeCun, Y., Huang, F.J., Bottou, L., 2004. Learning methods for generic object recognition with invariance to pose and lighting, in: Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on. IEEE, pp. II–97.

LeCun, Y., Kavukcuoglu, K., Farabet, C., 2010. Convolutional networks and applications in vision. Presented at the Proceedings of 2010 IEEE International Symposium on Circuits and Systems (ISCAS), pp. 253 –256.

Lee, B.B., 1996. Receptive field structure in the primate retina. Vision Res. 36, 631–644.

Lee, D.J., Gomez-Marin, O., Lam, B.L., 2000. Current Depression, Lifetime History of Depression, and Visual Acuity in Hispanic Adults. J. Vis. Impair. Blind. 94.

Leh, S.E., Chakravarty, M.M., Ptito, A., 2007. The Connectivity of the Human Pulvinar: A Diffusion Tensor Imaging Tractography Study. Int. J. Biomed. Imaging 2008.

Lejsek, H., Ásmundsson, F.H., Jónsson, B.T., Amsaleg, L., 2006. Scalability of local image descriptors: a comparative study. ACM, Santa Barbara, CA, USA, pp. 589–598.

Le, Q.V., Isbell, L.A., Matsumoto, J., Nguyen, M., Hori, E., Maior, R.S., Tomaz, C., Tran, A.H., Ono, T., Nishijo, H., 2013. Pulvinar neurons reveal neurobiological evidence of past selection for rapid detection of snakes. Proc. Natl. Acad. Sci. 201312648.

Le, Q.V., Monga, R., Devin, M., Corrado, G., Chen, K., Ranzato, M.A., Dean, J., Ng, A.Y., 2011. Building high-level features using large scale unsupervised learning. ArXiv Prepr. ArXiv11126209.

Li, J., Allinson, N.M., 2008. A comprehensive review of current local features for computer vision. Neurocomputing 71, 1771–1787.

Lin, C.-J., Chang, C.-C., 2001. LIBSVM A library for support vector machines. Softw. Available Httpwwwcsientuedutwcjlinlibsvm.

Liu, F., Picard, R.W., 1996. Periodicity, directionality, and randomness: Wold features for image modeling and retrieval. Pattern Anal. Mach. Intell. IEEE Trans. On 18, 722–733.

Liu, J., Liu, J., Xu, L., Jin, W., 2010. Electronic travel aids for the blind based on sensory substitution, in: 2010 5th International Conference on Computer Science and Education (ICCSE). Presented at the 2010 5th International Conference on Computer Science and Education (ICCSE), pp. 1328–1331.

Liu, X., 2008. A camera phone based currency reader for the visually impaired, in: Proceedings of the 10th International ACM SIGACCESS Conference on Computers and Accessibility. pp. 305–306.

Livingstone, M., Hubel, D., 1988. Segregation of form, color, movement, and depth: anatomy, physiology, and perception. Science 240, 740–749.

Li, Y., McCarthy, C., Barnes, N., 2012. On just noticeable difference for bionic eye, in: Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE. IEEE, pp. 2961–2964.

Loomis, J.M., 1974. Tactile letter recognition under different modes of stimulus presentation. Percept. Psychophys. 16, 401–408.

Loomis, J.M., 1985. Digital map and navigation system for the visually impaired (Unpublished manuscript). Department of Psychology, University of California Santa Barbara.

Loomis, J.M., Golledge, R.G., Klatzky, R.L., Speigle, J.M., Tietz, J., 1994. Personal guidance system for the visually impaired. pp. 85–91.

Loomis, J.M., Klatzky, R.L., Golledge, R.G., 2001. Navigating without vision: basic and applied research. Optom. Vis. Sci. 78, 282–289.

Loomis, J.M., Marston, J.R., Golledge, R.G., Klatzky, R.L., 2005. Personal guidance system for people with visual impairment: A comparison of spatial displays for route guidance. J. Vis. Impair. Blind. 99, 219–232.

Lowe, D.G., 2004. Distinctive Image Features from Scale-Invariant Keypoints. Int. J. Comput. Vis. 60, 91–110.

Lui, W.L.D., Browne, D., Kleeman, L., Drummond, T., Li, W.H., 2012. Transformative reality: Improving bionic vision with robotic sensing, in: Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE. IEEE, pp. 304–307.

Lu, X., Manduchi, R., 2005. Detection and localization of curbs and stairways using stereo vision, in: Robotics and Automation, 2005. ICRA 2005. Proceedings of the 2005 IEEE International Conference on. IEEE, pp. 4648–4654.

Macé, M.J.-M., Dramas, F., Jouffrais, C., 2012. Reaching to sound accuracy in the peri-personal space of blind and sighted humans, in: Computers Helping People with Special Needs. Springer, pp. 636–643.

Macskassy, S., Provost, F., 2004. Confidence bands for ROC curves: Methods and an empirical study.

Maidenbaum, S., Abboud, S., Amedi, A., 2014. Sensory substitution: Closing the gap between basic research and widespread practical visual rehabilitation. Neurosci. Biobehav. Rev., Multisensory integration, sensory substitution and visual rehabilitation 41, 3–15.

Maliene, V., Grigonis, V., Palevičius, V., Griffiths, S., 2011. Geographic information system: Old principles with new capabilities. Urban Des. Int. 16, 1–6.

Mallat, S.G., Zhang, Z., 1993. Matching pursuits with time-frequency dictionaries. Signal Process. IEEE Trans. On 41, 3397–3415.

Manduchi, R., Kurniawan, S., 2011. Mobility-related accidents experienced by people with visual impairment. Res. Pract. Vis. Impair. Blind. 4, 44–54.

Manjunath, B.S., Ohm, J.R., Vasudevan, V.V., Yamada, A., 2001. Color and texture descriptors. IEEE Trans. Circuits Syst. Video Technol. 11, 703–715.

Mann, H.B., Whitney, D.R., 1947. On a test of whether one of two random variables is stochastically larger than the other. Ann. Math. Stat. 18, 50–60.

Marg, E., Dierssen, G., 1965. Reported visual percepts from stimulation of the human brain with microelectrodes during therapeutic surgery. Stereotact. Funct. Neurosurg. 26, 57–75.

Margrain, T., 2000. Helping blind and partially sighted people to read: the effectiveness of low vision aids. Br. J. Ophthalmol. 84, 919–921.

Marr, D., 1982. Vision: A computational investigation into the human representation and processing of visual information, Henry Holt and Co. Inc N. Y. NY.

Marr, D., Hildreth, E., 1980. Theory of edge detection. Proc. R. Soc. Lond. B Biol. Sci. 207, 187–217.

Marr, D., Poggio, T., 1976. Cooperative computation of stereo disparity. Science 194, 283–287.

Marsala, C., Detyniecki, M., 2005. University of Paris 6 at TRECVID 2005: High-level feature extraction. TREC Video Retr. Eval. Online Proc. Novemb.

Marsala, C., Detyniecki, M., 2006. University of Paris 6 at TRECVID 2006: Forests of fuzzy decision trees for high-level feature extraction. TREC Video Retr. Eval. Online Proc. Novemb.

Martin, A., Doddington, G., Kamm, T., Ordowski, M., Przybocki, M., 1997. The DET curve in assessment of detection task performance. DTIC Document.

Martinez, S., Manuel, J., Ruiz, E., others, 2008. Stereo-based Aerial Obstacle Detection for the Visually Impaired. Presented at the Computer Vision Applications for the Visually Impaired (CVAVI 2008).

Martin, P.R., White, A.J., Goodchild, A.K., Wilder, H.D., Sefton, A.E., 1997. Evidence that Blue-on Cells are Part of the Third Geniculocortical Pathway in Primates. Eur. J. Neurosci. 9, 1536–1541.

Masland, R.H., 2001. The fundamental plan of the retina. Nat. Neurosci. 4, 877–886.

Mason, S.J., Graham, N.E., 2002. Areas beneath the relative operating characteristics (roc) and relative operating levels (rol) curves: Statistical significance and interpretation. Q. J. R. Meteorol. Soc. 128, 2145–2166.

Mattar, M.A., Hanson, A.R., Learned-Miller, E.G., 2005. Sign Classification using Local and Meta-Features, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, 2005. CVPR Workshops. Presented at the IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, 2005. CVPR Workshops, pp. 26–26.

Mau, S., Melchior, N., Makatchev, M., Steinfeld, A., 2008. BlindAid: An Electronic Travel Aid for the Blind. The Robotics Institute of Carnegie Mellon University.

Mayerhofer, B., Pressl, B., Wieser, M., 2008. ODILIA-A Mobility Concept for the Visually Impaired. Comput. Help. People Spec. Needs 1109–1116.

McCarthy, C., Barnes, N., 2012. Time-to-contact maps for navigation with a low resolution visual prosthesis, in: Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE. IEEE, pp. 2780–2783.

McCarthy, C., Barnes, N., Lieby, P., 2011. Ground surface segmentation for navigation with a low resolution visual prosthesis, in: Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE. IEEE, pp. 4457–4460.

McNeil, B.J., Hanley, J.A., 1983. Statistical approaches to the analysis of receiver operating characteristic (ROC) curves. Med. Decis. Mak. Int. J. Soc. Med. Decis. Mak. 4, 137–150.

Mehrotra, R., Gary, J.E., 1995. Similar-shape retrieval in shape data management. Computer 28, 57–62.

Mehtre, B.M., Kankanhalli, M.S., Lee, W.F., 1997. Shape measures for content based image retrieval: a comparison. Inf. Process. Manag. 33, 319–337.

Meijer, P.B., 1992. An experimental system for auditory image representations. Biomed. Eng. IEEE Trans. On 39, 112–121.

Merabet, L.B., Rizzo, J.F., Amedi, A., Somers, D.C., Pascual-Leone, A., 2005. What blindness can tell us about seeing again: merging neuroplasticity and neuroprostheses. Nat. Rev. Neurosci. 6, 71–77.

Metz, C.E., 1978. Basic principles of ROC analysis, in: Seminars in Nuclear Medicine. pp. 283–298.

Metz, C.E., Wang, P.-L., Kronman, H.B., 1984. A New Approach for Testing the Significance of Differences Between ROC Curves Measured from Correlated Data, in: Deconinck, F. (Ed.), Information Processing in Medical Imaging. Springer Netherlands, pp. 432–445.

Mikolajczyk, K., Schmid, C., 2004. Scale & Affine Invariant Interest Point Detectors. Int. J. Comput. Vis. 60, 63–86.

Mikolajczyk, K., Schmid, C., 2005. A performance evaluation of local descriptors. Pattern Anal. Mach. Intell. IEEE Trans. On 27, 1615–1630.

Minard, A., Misdariis, N., Houix, O., Susini, P., 2010. Catégorisation de sons environnementaux sur la base de profils morphologiques. 10ème Congrès Franccais Acoust.

Mindru, F., Tuytelaars, T., Gool, L.V., Moons, T., 2004. Moment invariants for recognition under changing viewpoint and illumination. Comput. Vis. Image Underst. 94, 3–27.

Mishkin, M., Ungerleider, L.G., Macko, K.A., 1983. Object vision and spatial vision: two cortical pathways. Trends Neurosci. 6, 414–417.

Mitchell, P., Hayes, P., Wang, J.J., 1997. Visual impairment in nursing home residents: the Blue Mountains Eye Study. Med. J. Aust. 166, 73–76.

Moller, H., Sorensen, M.F., Hammershoi, D., Jensen, C.B., 1995. Head-related transfer functions of human subjects. J. Audio Eng. Soc. 43, 300–321.

Moreels, P., Perona, P., 2007. Evaluation of Features Detectors and Descriptors based on 3D Objects. Int. J. Comput. Vis. 73, 263–284.

Mormiche, P., Boissonnat, V., 2003. Handicap et inégalités sociales : premiers apports de l'enquête « Handicaps, incapacités, dépendance ». Rev. Fr. Aff. Soc. n° 1-2, 267–285.

Mulckhuyse, M., Theeuwes, J., 2010. Unconscious attentional orienting to exogenous cues: A review of the literature. Acta Psychol. (Amst.) 134, 299–309.

Mumford, D., 1991. Mathematical theories of shape: Do they model perception?, in: San Diego,'91, San Diego, CA. pp. 2–10.

Mundy, J.L., 2006. Object recognition in the geometric era: A retrospective, in: Toward Category-Level Object Recognition. Springer, pp. 3–28.

Murase, H., Nayar, S.K., 1995. Visual learning and recognition of 3-D objects from appearance. Int. J. Comput. Vis. 14, 5–24.

Mussa-Ivaldi, F.A., Miller, L.E., 2003. Brain–machine interfaces: computational demands and clinical needs meet basic neuroscience. TRENDS Neurosci. 26, 329–334.

Naphade, M.R., Huang, T.S., 2000. A Probabilistic Framework for Semantic Indexing and Retrieval in Video. pp. 475–478.

Nashold Jr, B.S., 1970. Phosphenes resulting from stimulation of the midbrain in man. Arch. Ophthalmol. 84, 433.

Neumann, J. von, 1958. The Computer and the Brain. Yale University Press, New Haven, CT, USA.

Niblack, C.W., Barber, R., Equitz, W., Flickner, M.D., Glasman, E.H., Petkovic, D., Yanker, P., Faloutsos, C., Taubin, G., 1993. QBIC project: querying images by content, using color, texture, and shape, in: IS&T/SPIE's Symposium on Electronic Imaging: Science and Technology. pp. 173–187.

Nie, M., Ren, J., Li, Z., Niu, J., Qiu, Y., Zhu, Y., Tong, S., 2009. SoundView: An auditory guidance system based on environment understanding for the visually impaired people, in: Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2009. EMBC 2009. Presented at the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2009. EMBC 2009, pp. 7240–7243.

Niparko, J.K., 2009. Cochlear implants: Principles and practices. Wolters Kluwer Health.

Noë, A., O'Regan, J.K., 2002. On the brain-basis of visual consciousness: a sensorimotor account. Vis. Mind Sel. Read. Philos. Percept. 567–598.

Noë, A., Thompson, E., 2002. Vision and mind: Selected readings in the philosophy of perception. MIT Press.

Normann, R.A., 2007. Technology insight: future neuroprosthetic therapies for disorders of the nervous system. Nat. Clin. Pract. Neurol. 3, 444–452.

Normann, R.A., Greger, B.A., House, P.A., Romero, S.F., Pelayo, F., Fernández, E., 2009. Toward the development of a cortically based visual neuroprosthesis. J. Neural Eng. 6, 035001–035001.

Normann, R.A., Maynard, E.M., Rousche, P.J., Warren, D.J., 1999. A neural interface for a cortical vision prosthesis. Vision Res. 39, 2577–87.

Observatoire Régional de la Santé des Pays de la Loire, 2000. Les besoins de prise en charge de la malvoyance des personnes adultes et âgées dans le grand ouest. Situation actuelle et propositions.

Ojala, T., Pietikainen, M., Maenpaa, T., 2000. Gray scale and rotation invariant texture classification with local binary patterns. Lect. Notes Comput. Sci. 1842, 404–420.

Ojala, T., Pietikainen, M., Maenpaa, T., 2001. A generalized local binary pattern operator for multiresolution gray scale and rotation invariant texture classification. Lect. Notes Comput. Sci. 2013, 397–406.

Oozeer, M., Veraart, C., Legat, V., Delbeke, J., 2005. Simulation of intra-orbital optic nerve electrical stimulation. Med. Biol. Eng. Comput. 43, 608–617.

O'Regan, J.K., Noë, A., 2001. A sensorimotor account of vision and visual consciousness. Behav. Brain Sci. 24, 939–972.

Organisation Mondiale de la Santé, 1980. Classification internationale des handicaps: déficiences, incapacités et désavantages. OMSCTNERHILes Ed. INSERM.

Park, D.K., Jeon, Y.S., Won, C.S., 2000. Efficient use of local edge histogram descriptor. ACM New York, NY, USA, pp. 51–54.

Parlouar, R., Dramas, F., Macé, M., Jouffrais, C., 2009. Assistive device for the blind based on object recognition : an application to identify currency bills. Presented at the ACM Conference on Computers and Accessibility (ASSETS 2009), Pittsburgh, USA, pp. 227–228.

Parseihian, G., 2012. Sonification binaurale pour l'aide à la navigation. Université Pierre et Marie Curie - Paris VI.

Parseihian, G., Katz, B.F.G., 2012. Morphocons: A New Sonification Concept Based on Morphological Earcons. J. Audio Eng. Soc. 60, 409–418.

Pascolini, D., Mariotti, S.P., 2012. Global estimates of visual impairment: 2010. Br. J. Ophthalmol. 96, 614–618.

Pascual-Leone, A., Torres, F., 1993. Plasticity of the sensorimotor cortex representation of the reading finger in Braille readers. Brain 116, 39–52.

Pass, G., Zabih, R., 1996. Histogram refinement for content-based image retrieval. pp. 96–102.

Penfield, W., Rasmussen, T., 1950. The cerebral cortex of man; a clinical study of localization of function. Macmillan, Oxford, England.

Pentland, A.P., 1984. Fractal-based description of natural scenes. IEEE Trans. Pattern Anal. Mach. Intell. 661–674.

Pérez Fornos, A., Sommerhalder, J., Pittard, A., Safran, A.B., Pelizzone, M., 2008. Simulation of artificial vision: IV. Visual information required to achieve simple pointing and manipulation tasks. Vision Res. 48, 1705–1718.

Perrinet, L., 2004. Feature detection using spikes: The greedy approach. J. Physiol.-Paris, Decoding and interfacing the brain: from neuronal assemblies to cyborgs 98, 530–539.

Perrinet, L., Samuelides, M., Thorpe, S., 2004. Sparse spike coding in an asynchronous feed-forward multi-layer neural network using matching pursuit. Neurocomputing, New Aspects in Neurocomputing: 10th European Symposium on Artificial Neural Networks 2002 57, 125–134.

Perrinet, L., Samuelides, M., Thorpe, S., 2004. Coding static natural images using spiking event times: do neurons Cooperate? IEEE Trans. Neural Netw. 15, 1164–1175.

Petrie, H., Johnson, V., Strothotte, T., Raab, A., Fritz, S., Michel, R., 1996. MoBIC: Designing a travel aid for blind and elderly people. J. Navig. 49, 45–52.

Petrie, H., Johnson, V., Strothotte, T., Raab, A., Michel, R., Reichert, L., Schalt, A., 1997. MoBIC: An Aid to Increase the Independent Mobility of Blind Travellers. Br. J. Vis. Impair. 15, 63–66.

Philbin, J., Marin-Jimenez, M., Srinivasan, S., Zisserman, A., Jain, M., Vempati, S., Sankar, P., Jawahar, C.V., 2008. Oxford/IIIT TRECVID 2008–Notebook paper.

Piccolino, M., 1995. The feedback synapse from horizontal cells to cone photoreceptors in the vertebrate retina. Prog. Retin. Eye Res. 14, 141–196.

Pissaloux, E., Velazquez, R., Maingreaud, F., 2008. Intelligent glasses: A multimodal interface for data communication to the visually impaired, in: IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems, 2008. MFI 2008. Presented at the IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems, 2008. MFI 2008, pp. 120–124.

Poggio, T., Edelman, S., 1990. A network that learns to recognize 3D objects. Nature 343, 263–266.

Powers, D.M.W., 2011. Evaluation: From Precision, Recall and F-Measure to ROC., Informedness, Markedness & Correlation. J. Mach. Learn. Technol. 2, 37–63.

Powers, D.M.W., 2012. The Problem of Area Under the Curve. pp. 567–573.

Pradeep, V., Medioni, G., Weiland, J., 2008. Piecewise Planar Modeling for Step Detection using Stereo Vision. Presented at the Computer Vision Applications for the Visually Impaired (CVAVI 2008).

Prokop, R.J., Reeves, A.P., 1992. A survey of moment-based techniques for unoccluded object representation and recognition. CVGIP Graph. Models Image Process. 54, 438–460.

Provost, F.J., Fawcett, T., 1997. Analysis and Visualization of Classifier Performance: Comparison under Imprecise Class and Cost Distributions., in: KDD. pp. 43–48.

Ptito, M., Fumal, A., de Noordhout, A.M., Schoenen, J., Gjedde, A., Kupers, R., 2008. TMS of the occipital cortex induces tactile sensations in the fingers of blind Braille readers. Exp. Brain Res. 184, 193–200.

Ptito, M., Kupers, R., 2005. Cross-modal plasticity in early blindness. J. Integr. Neurosci. 4, 479–488.

Ptito, M., Matteau, I., Gjedde, A., Kupers, R., 2009. Recruitment of the middle temporal area by tactile motion in congenital blindness: NeuroReport 20, 543–547.

Ptito, M., Moesgaard, S.M., Gjedde, A., Kupers, R., 2005. Cross-modal plasticity revealed by electrotactile stimulation of the tongue in the congenitally blind. Brain 128, 606–614.

Rahman, S., Shah, A.S.M.Z., Whitney, G., 2004. Computer Vision Based Navigation System for the Visually Impaired, in: ACM SIGGRAPH 2004 Posters, SIGGRAPH '04. ACM, New York, NY, USA, p. 64–.

Ran, L., Helal, S., Moore, S., 2004. Drishti: an integrated indoor/outdoor blind navigation system and service.

Resnikoff, S., Pascolini, D., Etya'ale, D., Kocur, I., Pararajasegaram, R., Pokharel, G.P., Mariotti, S.P., 2004. Global data on visual impairment in the year 2002. Bull. World Health Organ. 82, 844–851.

Resnikoff, S., Pascolini, D., Mariottia, S.P., Pokharela, G.P., 2008. Global magnitude of visual impairment caused by uncorrected refractive errors in 2004. Bull. World Health Organ. 86, 63–70.

Ribeiro, F., Florêncio, D., Chou, P.A., Zhang, Z., 2012. Auditory augmented reality: Object sonification for the visually impaired, in: Multimedia Signal Processing (MMSP), 2012 IEEE 14th International Workshop on. IEEE, pp. 319–324.

Richefeu, J., 2006. Détection et analyse du mouvement sur système de vision à base de rétine numérique. ENSTA ParisTech.

Riesenhuber, M., Poggio, T., 2002. Neural mechanisms of object recognition. Curr. Opin. Neurobiol. 12, 162–168.

Rigoulot, S., 2008. Impact comportemental et électrophysiologique de l'information émotionnelle en vision périphérique. Université du Droit et de la Santé - Lille II.

Ristic, B., Arulampalm, S., Gordon, N.J., 2004. Beyond the Kalman filter: Particle filters for tracking applications. Artech House Publishers.

Rizzo, J.F., Wyatt, J., 1997. Prospects for a Visual Prosthesis. The Neuroscientist 3, 251–262.

Rizzo, S., Belting, C., Cinelli, L., Allegrini, L., Genovesi-Ebert, F., Barca, F., di Bartolo, E., 2014. The Argus II Retinal Prosthesis: Twelve-Month Outcomes from a Single-Study Center. Am. J. Ophthalmol.

Roberts, L.G., 1963. Machine Perception of Three-Dimensional Soups. Massachusetts Institute of Technology.

Robinson, D.L., Cowie, R.J., 1997. The primate pulvinar: structural, functional, and behavioral components of visual salience. Thalamus 2, 53–92.

Rodieck, R.W., 1998. The first steps in seeing. Sinauer Associates Sunderland, MA.

Roentgen, U.R., Gelderblom, G.J., Soede, M., de Witte, L.P., 2008. Inventory of Electronic Mobility Aids for Persons with Visual Impairments: A Literature Review. J. Vis. Impair. Blind. 102, 702–724.

Ross, D.A., Blasch, B.B., 2000. Wearable interfaces for orientation and wayfinding, in: Proceedings of the Fourth International ACM Conference on Assistive Technologies. ACM, pp. 193–200.

Roth, P.M., Winter, M., 2008. Survey of appearance-based methods for object recognition. Inst Comput. Graph. Vis. Graz Univ. Technol. Austria Tech. Rep. ICGTR0108 ICG-TR-0108.

Rowley, H.A., Baluja, S., Kanade, T., 1998. Neural network-based face detection, in: IEEE Transactions on Pattern Analysis and Machine Intelligence. pp. 23–38.

Rq, I., Rg, C., P, M., K, A., 1998. Visual impairment and falls in older adults: the Blue Mountains Eye Study. J. Am. Geriatr. Soc. 46, 58–64.

Rubin, G.S., Bandeen-Roche, K., Huang, G.-H., Munoz, B., Schein, O.D., Fried, L.P., West, S.K., 2001. The Association of Multiple Visual Impairments with Self-Reported Visual Disability: SEE Project. Invest Ophthalmol Vis Sci 42, 64–72.

Rui, Y., Huang, T.S., Chang, S.F., 1997. Image retrieval: Past, present, and future.

Saalmann, Y.B., Pinsk, M.A., Wang, L., Li, X., Kastner, S., 2012. The Pulvinar Regulates Information Transmission Between Cortical Areas Based on Attention Demands. Science 337, 753–756.

Salive, M.E., Guralnik, J., Christen, W., Glynn, R.J., Colsher, P., Ostfeld, A.M., 1992. Functional Blindness and Visual Impairment in Older Adults from Three Communities. Ophthalmology 99, 1840–1847.

Salive, M.E., Guralnik, J., Glynn, R.J., Christen, W., Wallace, R.B., Ostfeld, A.M., 1994. Association of visual impairment with mobility and physical function. J. Am. Geriatr. Soc. 42, 287–292.

Sampaio, E., Maris, S., Bach-y-Rita, P., 2001. Brain plasticity:"visual"acuity of blind persons via the tongue. Brain Res. 908, 204–207.

Sander, M.-S., Lelièvre, F., Tallec, A., Bournot, M.-C., 2005. Les personnes ayant un handicap visuel, Les apports de l'enquête Handicaps - Incapacités - Dépendance, Rapport d'enquete du Ministere Francais de la Santé et des Solidarités. Ministere Francais de la Santé et des Solidarités.

Santos A, Humayun MS, de Juan E, Jr, et al, 1997. Preservation of the inner retina in retinitis pigmentosa: A morphometric analysis. Arch. Ophthalmol. 115, 511–515.

Scassellati, B.M., Alexopoulos, S., Flickner, M.D., 1994. Retrieving images by 2D shape: a comparison of computation methods with human perceptual judgments, in: IS&T/SPIE 1994 International Symposium on Electronic Imaging: Science and Technology. pp. 2–14.

Scharstein, D., Szeliski, R., 2002. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. Int. J. Comput. Vis. 47, 7–42.

Schmid, C., Mohr, R., 1997. Local grayvalue invariants for image retrieval. Pattern Anal. Mach. Intell. IEEE Trans. On 19, 530–535.

Schmidt, E.M., Bak, M.J., Hambrecht, F.T., Kufta, C.V., O'Rourke, D.K., Vallabhanath, P., 1996. Feasibility of a visual prosthesis for the blind based on intracortical micro stimulation of the visual cortex. Brain 119, 507–522.

Sclaroff, S., Pentland, A.P., 1995. Modal matching for correspondence and recognition. Pattern Anal. Mach. Intell. IEEE Trans. On 17, 545–561.

Sebastian, T.B., Klein, P.N., Kimia, B.B., 2001. Recognition of Shapes by Editing Shock Graphs., in: ICCV. pp. 755–762.

Serre, T., Wolf, L., Poggio, T., 2005. Object recognition with features inspired by visual cortex. pp. 994–1000.

Se, S., Brady, M., 1997. Vision-based Detection of Kerbs and Steps., in: BMVC.

Sewards, T.V., Sewards, M.A., 2002. Innate visual object recognition in vertebrates: some proposed pathways and mechanisms. Comp. Biochem. Physiol. A. Mol. Integr. Physiol. 132, 861–891.

Shafi, S., Hassan, S.M., Arshaq, A., Khan, M.J., Shamail, S., 2008. Software quality prediction techniques: A comparative analysis, in: 4th International Conference on Emerging Technologies, 2008. ICET 2008. Presented at the 4th International Conference on Emerging Technologies, 2008. ICET 2008, pp. 242–246.

Sheskin, D.J., 2000. Parametric and nonparametric statistical procedures. Boca Raton CRC.

Sieber, R., 2006. Public participation geographic information systems: A literature review and framework. Ann. Assoc. Am. Geogr. 96, 491–507.

Silapachote, P., Weinman, J., Hanson, A., Mattar, M.A., Weiss, R., 2005. Automatic sign detection and recognition in natural scenes, in: Computer Vision and Pattern Recognition-Workshops, 2005. CVPR Workshops. IEEE Computer Society Conference on. IEEE, pp. 27–27.

Skulimowski, P., Strumillo, P., 2008. Refinement of depth from stereo camera ego-motion parameters. Electron. Lett. 44, 729–730.

Smeaton, A.F., Over, P., Kraaij, W., 2008. High-Level Feature Detection from Video in TRECVid: a 5-Year Retrospective of Achievements. Multimed. Content Anal. Theory Appl. Ed.

Smith, J.R., Chang, S.F., 1996a. Tools and techniques for color image retrieval. Storage Retr. Image Video Databases IV 2670, 426–437.

Smith, J.R., Chang, S.F., 1996b. Automated binary texture feature sets for image retrieval.

Solomon, S.G., Lennie, P., 2007. The machinery of colour vision. Nat. Rev. Neurosci. 8, 276–286.

Souza, C.R., 2010. Kernel functions for machine learning applications [WWW Document]. URL http://crsouza.blogspot.com/2010/03/kernel-functions-for-machine-learning.html

Sterling, P., 2004. How Retinal Circuits Optimize the Transfer of Visual Information, in: The Visual Neurosciences.

Stirling, R., Collin, J., Fyfe, K., Lachapelle, G., 2003. An innovative shoe-mounted pedestrian navigation system, in: Proceedings of European Navigation Conference GNSS. pp. 110–5.

Stockman, A., Sharpe, L.T., 2006. Into the twilight zone: the complexities of mesopic vision and luminous efficiency. Ophthalmic Physiol. Opt. 26, 225–239.

Stone JL, Barlow WE, Humayun MS, de Juan E, Jr, Milam AH, 1992. Morphometric analysis of macular photoreceptors and ganglion cells in retinas with retinitis pigmentosa. Arch. Ophthalmol. 110, 1634–1639.

Stricker, M., Dimai, A., 1996. Color indexing with weak spatial constraints. Storage Retr. Image Video Databases IV 2670.

Stricker, M., Orengo, M., 1995. Similarity of color images. San Jose CA USA, pp. 381–392.

Striem-Amit, E., Guendelman, M., Amedi, A., 2012. "Visual"Acuity of the Congenitally Blind Using Visual-to-Auditory Sensory Substitution. PloS One 7, e33136.

Strothotte, T., Fritz, S., Michel, R., Raab, A., Petrie, H., Johnson, V., Reichert, L., Schalt, A., 1996. Development of dialogue systems for a mobility aid for blind people: initial design and usability testing, in: Proceedings of the Second Annual ACM Conference on Assistive Technologies. ACM, pp. 139–144.

Sui, X., Li, L., Chai, X., Wu, K., Zhou, C., Sun, X., Xu, X., Li, X., Ren, Q., 2009. Visual Prosthesis for Optic Nerve Stimulation, in: Greenbaum, E., Zhou, D. (Eds.), Implantable Neural Prostheses 1. Springer US, New York, NY, pp. 43–83.

Swain, M.J., Ballard, D.H., 1990. Indexing via color histograms. pp. 390–393.

Szeto, A.Y. j, Saunders, F.A., 1982. Electrocutaneous Stimulation for Sensory Communication in Rehabilitation Engineering. IEEE Trans. Biomed. Eng. BME-29, 300–308.

Takizawa, H., Yamaguchi, S., Aoyagi, M., Ezaki, N., Mizuno, S., 2012. Kinect cane: An assistive system for the visually impaired based on three-dimensional object recognition, in: 2012 IEEE/SICE International Symposium on System Integration (SII). Presented at the 2012 IEEE/SICE International Symposium on System Integration (SII), pp. 740–745.

Tamura, H., Mori, S., Yamawaki, T., 1978. Textural Features Corresponding to Visual Perception. IEEE Trans. Syst. Man Cybern. 8, 460–473.

Tang, H., Beebe, D.J., 2003. Design and microfabrication of a flexible oral electrotactile display. Microelectromechanical Syst. J. Of 12, 29–36.

Tang, H., Beebe, D.J., 2006. An oral tactile interface for blind navigation. Neural Syst. Rehabil. Eng. IEEE Trans. On 14, 116–123.

Tarel, J.-P., Gagalowicz, A., 1995. Calibration de caméra à base d'ellipses. Trait. Signal 12, 177–187.

Taylor, H.R., Keeffe, J.E., Vu, H.T., Wang, J.J., Rochtchina, E., Pezzullo, M.L., Mitchell, P., 2005. Vision loss in Australia. Med J Aust 182, 565–568.

Tekin, E., Coughlan, J.M., 2010. A mobile phone application enabling visually impaired users to find and read product barcodes, in: Computers Helping People with Special Needs. Springer, pp. 290–295.

The Eye Diseases Prevalence Research Group, 2004. Causes and prevalence of visual impairment among adults in the unitedstates. Arch. Ophthalmol. 122, 477–485.

Thompson, D., Mundy, J.L., 1987. Three-dimensional model matching from an unconstrained viewpoint, in: 1987 IEEE International Conference on Robotics and Automation. Proceedings. Presented at the 1987 IEEE International Conference on Robotics and Automation. Proceedings, pp. 208–220.

Thorpe, S., Delorme, A., Van Rullen, R., 2001. Spike-based strategies for rapid processing. Neural Netw. 14, 715–725.

Thorpe, S., Fize, D., Marlot, C., 1996. Speed of processing in the human visual system. nature 381, 520–522.

Thorpe, S.J., Delorme, A., Van Rullen, R., Paquier, W., 2000. Reverse engineering of the visual system using networks of spiking neurons. Presented at the Circuits and Systems, 2000. Proceedings. ISCAS 2000 Geneva. The 2000 IEEE International Symposium on, pp. 405–408 vol.4.

Thorpe, S.J., Guyonneau, R., Guilbaud, N., Allegraud, J.M., VanRullen, R., 2004. SpikeNet: Real-time visual processing with one spike per neuron. Neurocomputing 58, 857–864.

Thorpe, S.J., Imbert, M., 1989. Biological constraints on connectionist modelling. Connect. Perspect. 63–92.

Thylefors, B., Negrel, A.D., Pararajasegaram, R., Dadzie, K.Y., 1995. Global data on blindness. Bull. World Health Organ. 73, 115–121.

Tielsch, J.M., Sommer, A., Witt, K., Katz, J., Royall, R.M., 1990. Blindness and visual impairment in an American urban population: the Baltimore Eye Survey. Arch. Ophthalmol. 108, 286–290.

Tirthapura, S., Sharvit, D., Klein, P., Kimia, B.B., 1998. Indexing based on edit-distance matching of shape graphs, in: Photonics East (ISAM, VVDC, IEMB). pp. 25–36.

Tombari, F., Mattoccia, S., Di Stefano, L., 2007. Segmentation-based adaptive support for accurate stereo correspondence, in: Advances in Image and Video Technology. Springer, pp. 427–438.

Tombari, F., Mattoccia, S., Di Stefano, L., Addimanda, E., 2008. Classification and evaluation of cost aggregation methods for stereo correspondence, in: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. pp. 1–8.

Troyk, P., Bak, M., Berg, J., Bradley, D., Cogan, S., Erickson, R., Kufta, C., McCreery, D., Schmidt, E., Towle, V., 2003. A model for intracortical visual prosthesis research. Artif. Organs 27, 1005–1015.

Tsai, D., Morley, J.W., Suaning, G.J., Lovell, N.H., 2009. A wearable real-time image processor for a vision prosthesis. Comput. Methods Programs Biomed. 95, 258–269.

Tsukada, K., Yasumura, M., 2004. Activebelt: Belt-type wearable tactile display for directional navigation, in: UbiComp 2004: Ubiquitous Computing. Springer, pp. 384–399.

Tuceryan, M., Jain, A.K., 1993. Texture analysis. Handb. Pattern Recognit. Comput. Vis. 235–276.

Turano, K.A., Broman, A.T., Bandeen-Roche, K., Munoz, B., Rubin, G.S., West, S.K., 2004. Association of visual field loss and mobility performance in older adults: Salisbury Eye Evaluation study. Optom. Vis. SCI 81, 298–307.

Turano, K.A., Rubin, G.S., Quigley, H.A., 1999. Mobility Performance in Glaucoma. Invest. Ophthalmol. Vis. Sci. 40, 2803–2809.

Ulrich, I., Borenstein, J., 2001. The GuideCane-applying mobile robot technologies to assist the visually impaired. Syst. Man Cybern. Part Syst. Hum. IEEE Trans. On 31, 131–136.

Ulrich, I., Nourbakhsh, I., 2000. Appearance-based obstacle detection with monocular color vision, in: AAAI/IAAI. pp. 866–871.

Valbuena, M., Bandeen-Roche, K., Rubin, G.S., Munoz, B., West, S.K., 1999. Self-reported assessment of visual function in a population-based study: the SEE project. Salisbury Eye Evaluation. Invest. Ophthalmol. Vis. Sci. 40, 280.

Van Der Heijden, F., Regtien, P.P.L., 2005. Wearable navigation assistance-a tool for the blind. Meas. Sci. Rev. 5, 53–56.

Van de Sande, K.E.A., Gevers, T., Snoek, C.G.M., 2008. Evaluation of color descriptors for object and scene recognition. pp. 1–8.

VanNewkirk, M.R., Weih, L., McCarty, C.A., Taylor, H.R., 2001. Cause-specific prevalence of bilateral visual impairment in Victoria, Australia: The visual impairment project. Ophthalmology 108, 960–967.

Van Otterloo, P.J., 1988. A contour-oriented approach to digital shape analysis. Technische Universiteit Delft.

VanRullen, R., Gautrais, J., Delorme, A., Thorpe, S., 1998. Face processing using one spike per neurone. Biosystems 48, 229–239.

VanRullen, R., Guyonneau, R., Thorpe, S.J., 2005. Spike times make sense. Trends Neurosci. 28, 1–4.

VanRullen, R., Thorpe, S.J., 2001. Rate coding versus temporal order coding: what the retinal ganglion cells tell the visual cortex. Neural Comput. 13, 1255–1283.

VanRullen, R., Thorpe, S.J., 2002. Surfing a spike wave down the ventral stream. Vision Res. 42, 2593–2615.

Vapnik, V.N., 1995. The nature of statistical learning theory. springer.

Varela, F.J., 1992. Whence perceptual meaning? A cartography of current ideas, in: Understanding Origins. Springer, pp. 235–263.

Varma, R., Wu, J., Chong, K., Azen, S., Hays, R., 2006. Impact of Severity and Bilaterality of Visual Impairment on Health-Related Quality of Life. Ophthalmology 113, 1846–1853.

Velázquez, R., 2010. Wearable assistive devices for the blind, in: Wearable and Autonomous Biomedical Devices and Systems for Smart Environment. Springer, pp. 331–349.

Velázquez, R., Pissaloux, E., Szewczyk, J., Hafez, M., 2004. Système visuo-tactile d'aide à la mobilité indépendante des déficients visuels dans des environnements 3D non - coopérants. J3eA 3.

Veltkamp, R.C., Hagedoorn, M., 2001. State of the art in shape matching. Springer.

Veltkamp, R.C., Tanase, M., 2002. Content-based image retrieval systems: A survey. Dep. Comput. Sci. Utrecht Univ. 1–62.

Veraart, C., 1989. Neurophysiological approach to the design of visual prostheses: a theoretical discussion. J. Med. Eng. Technol. 13, 57–62.

Veraart, C., Duret, F., Brelén, M., Oozeer, M., Delbeke, J., 2004. Vision rehabilitation in the case of blindness. Expert Rev. Med. Devices 1, 139–153.

Veraart, C., Raftopoulos, C., Mortimer, J.T., Delbeke, J., Pins, D., Michaux, G., Vanlierde, A., Parrini, S., Wanet-Defalque, M.-C., 1998. Visual sensations produced by optic nerve stimulation using an implanted self-sizing spiral cuff electrode. Brain Res. 813, 181–186.

Veraart, C., Wanet-Defalque, M.-C., Gérard, B., Vanlierde, A., Delbeke, J., 2003. Pattern Recognition with the Optic Nerve Visual Prosthesis. Artif. Organs 27, 996–1004.

Viitaniemi, V., Laaksonen, J., 2006. Techniques for still image scene classification and object detection, in: Artificial Neural Networks–ICANN 2006. Springer, pp. 35–44.

Völkel, T., Weber, G., 2007. A New Approach for Pedestrian Navigation for Mobility Impaired Users Based on Multimodal Annotation of Geographical Data, in: Stephanidis, C. (Ed.), Universal Access in Human-Computer Interaction. Ambient Interaction, Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 575–584.

Vuilleumier, P., Armony, J.L., Driver, J., Dolan, R.J., 2003. Distinct spatial frequency sensitivities for processing faces and emotional expressions. Nat. Neurosci. 6, 624–631.

Wahl, H.-W., Schilling, O., Oswald, F., Heyl, V., 1999. Psychosocial consequences of age-related visual impairment: Comparison with mobility-impaired older adults and long-term outcome. J. Gerontol. B. Psychol. Sci. Soc. Sci. 54, P304–P316.

Walker, B.N., Lindsay, J., 2005. Using virtual reality to prototype auditory navigation displays. Assist. Technol. J. 17, 72–81.

Walker, B.N., Lindsay, J., 2005. Navigation performance in a virtual environment with bonephones, in: Proceedings of the International Conference on Auditory Display (ICAD2005). pp. 1–26.

Walker, B.N., Lindsay, J., 2006. Navigation performance with a virtual auditory display: Effects of beacon sound, capture radius, and practice. Hum. Factors J. Hum. Factors Ergon. Soc. 48, 265–278.

Wallhagen, M.I., Strawbridge, W.J., Shema, S.J., Kurata, J., Kaplan, G.A., 2001. Comparative impact of hearing and vision impairment on subsequent functioning. J. Am. Geriatr. Soc. 49, 1086–1092.

Wandell, B.A., 1995. Foundations of vision. Sinauer Associates.

Wang, J.J., Mitchell, P., Cumming, R.G., Smith, W., 2003. Visual impairment and nursing home placement in older Australians: the Blue Mountains Eye Study. Ophthalmic Epidemiol. 10, 3–13.

Wang, J.J., Mitchell, P., Simpson, J.M., Cumming, R.G., Smith, W., 2001. Visual impairment, age-related cataract, and mortality. Arch. Ophthalmol. 119, 1186–1190.

Wang, J.J., Mitchell, P., Smith, W., 2000. Vision and Low Self-Rated Health: The Blue Mountains Eye Study. Invest. Ophthalmol. Vis. Sci. 41.

Wang, J.J., Mitchell, P., Smith, W., Cumming, R.G., Attebo, K., 1999. Impact of visual impairment on use of community support services by elderly persons: the Blue Mountains Eye Study. Invest. Ophthalmol. Vis. Sci. 40, 12–19.

Wang, Z.-F., Zheng, Z.-G., 2008. A region based stereo matching algorithm using cooperative optimization, in: IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008. Presented at the IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008, pp. 1–8.

Ward, J., Meijer, P., 2010. Visual experiences in the blind induced by an auditory sensory substitution device. Conscious. Cogn. 19, 492–500.

Ward, R., Danziger, S., Bamford, S., 2005. Response to Visual Threat Following Damage to the Pulvinar. Curr. Biol. 15, 571–573.

Wässle, H., 2004. Parallel processing in the mammalian retina. Nat. Rev. Neurosci. 5, 747–757.

West, S.K., Munoz, B., Rubin, G.S., Schein, O.D., Bandeen-Roche, K., Zeger, S., German, S., Fried, L.P., 1997. Function and visual impairment in a population-based study of older adults. The SEE project. Salisbury Eye Evaluation. Invest. Ophthalmol. Vis. Sci. 38, 72.

West, S.K., Rubin, G.S., Broman, A.T., Munoz, B., Bandeen-Roche, K., Turano, K., 2002. How does visual impairment affect performance on tasks of everyday life?: The SEE Project. Arch. Ophthalmol. 120, 774.

Whitson, H.E., Cousins, S.W., Burchett, B.M., Hybels, C.F., Pieper, C.F., Cohen, H.J., 2007. The Combined Effect of Visual Impairment and Cognitive Impairment on Disability in Older People. J. Am. Geriatr. Soc. 55, 885–891.

Willis, S., Helal, S., 2005. RFID Information Grid for Blind Navigation and Wayfinding., in: ISWC. pp. 34–37.

Wilson, B.S., Dorman, M.F., 2008. Cochlear implants: a remarkable past and a brilliant future. Hear. Res. 242, 3–21.

Wilson, J., Walker, B.N., Lindsay, J., Cambias, C., Dellaert, F., 2007. Swan: System for wearable audio navigation.

Wirth, M., Fraschini, M., Masek, M., Bruynooghe, M., 2006. Performance evaluation in image processing. EURASIP J. Appl. Signal Process. 2006, 211–211.

Worchel, P., Dallenbach, K.M., 1947. Facial Vision: Perception of Obstacles by the Deaf-Blind. Am. J. Psychol. 60, 502–553.

World Health Organization, 2005. State of the world's sight : VISION 2020 : the Right to Sight : 1999-2005. Geneva : World Health Organization.

World Health Organization, 2010a. Action plan for the prevention of avoidable blindness and visual impairment, 2009-2013.

World Health Organization, 2010b. Global data on visual impairments : 2010.

Wormald, R.P., Wright, L.A., Courtney, P., Beaumont, B., Haines, A.P., 1992. Visual problems in the elderly population and implications for services. BMJ 304, 1226.

Würtz, R.P., Lourens, T., 1997. Corner detection in color images by multiscale combination of end-stopped cortical cells, in: Gerstner, W., Germond, A., Hasler, M., Nicoud, J.-D. (Eds.), Artificial Neural Networks — ICANN'97, Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 901–906.

Würtz, R.P., Lourens, T., 2000. Corner detection in color images through a multiscale combination of end-stopped cortical cells. Image Vis. Comput. 18, 531–541.

Xie, D., Yan, T., Ganesan, D., Hanson, A., 2008. Design and implementation of a dual-camera wireless sensor network for object retrieval, in: Proceedings of the 7th International Conference on Information Processing in Sensor Networks. IEEE Computer Society, pp. 469–480.

Yang, J., Jiang, Y.G., Hauptmann, A.G., Ngo, C.W., 2007. Evaluating bag-of-visual-words representations in scene classification. ACM New York, NY, USA, pp. 197–206.

Yang, Q., Wang, L., Yang, R., Stewenius, H., Nister, D., 2009. Stereo Matching with Color-Weighted Correlation, Hierarchical Belief Propagation, and Occlusion Handling. IEEE Trans. Pattern Anal. Mach. Intell. 31, 492–504.

Yan Ke, Sukthankar, R., 2004. PCA-SIFT: a more distinctive representation for local image descriptors. Presented at the Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on, pp. II–506–II–513 Vol.2.

Yan, L., Pan, Z., Winstanley, A.C., Fotheringham, A.S., Zheng, J., 2009. Feedback Control Models and Their Application in Pedestrian Navigation Systems.

Yates, D., 2012. Attention: The focus of attention. Nat. Rev. Neurosci. 13, 666–666.

Yi, C., Flores, R.W., Chincha, R., Tian, Y., 2013. Finding objects for assisting blind people. Netw. Model. Anal. Health Inform. Bioinforma. 1–9.

Yoon, K.-J., Kweon, I.S., 2006. Adaptive support-weight approach for correspondence search. Pattern Anal. Mach. Intell. IEEE Trans. On 28, 650–656.

Zagler, W.L., Mayer, P., Winkler, N., Busboom, M., 1992. Microprocessor devices to lower the barriers for the blind and visually impaired. J. Microcomput. Appl. 15, 57–64.

Zelek, J., Audette, R., Balthazaar, J., Dunk, C., 1999. A stereo-vision system for the visually impaired. Sch. Eng. Univ. Guelph.

Zhan, F.B., 1997. Three fastest shortest path algorithms on real road networks: Data structures and procedures. J. Geogr. Inf. Decis. Anal. 1, 69–82.

Zhan, F.B., Noon, C.E., 1998. Shortest path algorithms: an evaluation using real road networks. Transp. Sci. 32, 65–73.

Zhang, D., Lu, G., 2004. Review of shape representation and description techniques. Pattern Recognit. 37, 1–19.

Zhang, Q., Tolias, G., Mansencal, B., Saracoglu, A., Aginako, N., Alatan, A., Alexandre, L.A., Avrithis, Y., Benois-Pineau, J., Chandramouli, K., 2008. COST292 experimental framework for TRECVID 2008.

Zhao, Y., Tian, Y., Liu, H., Ren, Q., Chai, X., 2008. Pixelized images recognition in simulated prosthetic vision, in: 7th Asian-Pacific Conference on Medical and Biological Engineering. Springer, pp. 492–496.

Zheng, J., Winstanley, A., Pan, Z., Coveney, S., 2009. Spatial Characteristics of Walking Areas for Pedestrian Navigation, in: Third International Conference on Multimedia and Ubiquitous Engineering, 2009. MUE '09. Presented at the Third International Conference on Multimedia and Ubiquitous Engineering, 2009. MUE '09, pp. 452–458.

Zhong, Y., Garrigues, P.J., Bigham, J.P., 2013. Real time object scanning using a mobile phone and cloud-based visual search engine, in: Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility. ACM, p. 20.

Zhu, J., Wang, L., Yang, R., Davis, J., 2008. Fusion of time-of-flight depth and stereo for high accuracy depth maps, in: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. pp. 1–8.

Zrenner, E., Miliczek, K.-D., Gabel, V.P., Graf, H.G., Guenther, E., Haemmerle, H., Hoefflinger, B., Kohler, K., Nisch, W., Schubert, M., Stett, A., Weiss, S., 1997. The Development of Subretinal Microphotodiodes for Replacement of Degenerated Photoreceptors. Ophthalmic Res. 29, 269–280.

Annexes

Sommaire de section

1. Organisation du système visuel humain

1.1 L'œil

L'œil constitue l'organe perceptif de la vision. Premier élément de la chaîne de traitement de l'information visuelle, il a pour fonctions principales de focaliser les rayons lumineux entrants afin de former une image sur la rétine, puis de transformer ces ondes électromagnétiques en influx nerveux qui seront transmis par les nerfs optiques dans les aires visuelles du cerveau.

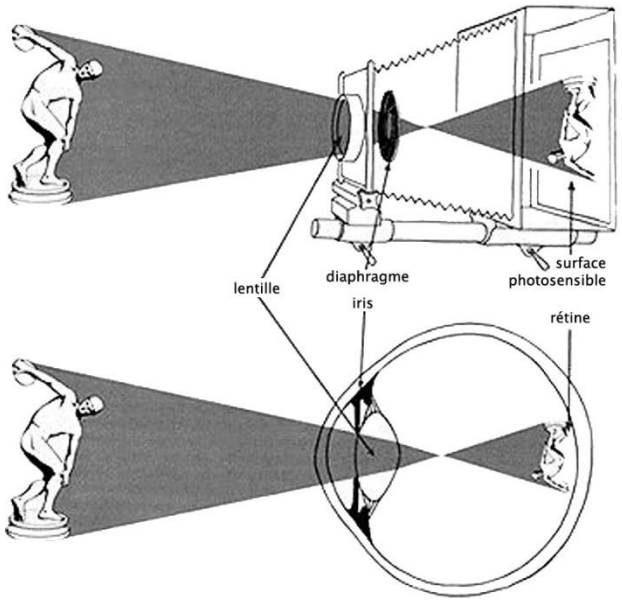


Figure 1 Analogies entre les structures optiques de l'œil et d'un appareil photo

A la manière d'un appareil photo, les différentes structures optiques de l'œil vont permettre de concentrer la lumière sur un plan focal. Pour cela, les rayons subiront plusieurs réfractions[1] en traversant successivement différents milieux transparents : la cornée, l'humeur aqueuse, le cristallin, et enfin l'humeur vitrée. D'un point de vue anatomique, la paroi du globe oculaire est constituée de trois enveloppes, aussi appelées tuniques, visibles dans la Figure 2.

[1] La réfraction est la déviation des rayons lumineux passant obliquement d'un milieu transparent à un autre. L'angle de réfraction dépend à la fois de l'angle d'incidence, et de l'indice de réfraction de chacun des 2 milieux

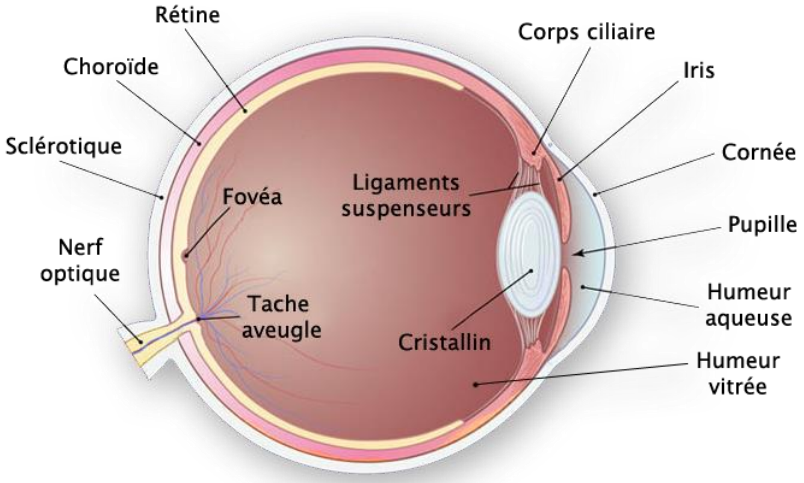


Figure 2 Anatomie de l'œil humain

La couche externe est constituée pour près de 80% par la sclérotique, une membrane blanche et opaque très résistante, permettant de protéger l'œil et de contenir sa pression interne. Dans sa partie antérieure, la sclère se prolonge par la cornée, une membrane plus fine et plus fragile mais transparente, laissant entrer les rayons lumineux. Celle-ci, remplie d'un liquide appelé l'humeur aqueuse, constitue la principale lentille optique de l'œil. Pour la protéger, elle est entourée d'une zone plus résistante, la conjonctive, qui recouvre également la face interne des paupières.

La tunique moyenne est composée de la choroïde, des corps ciliaires et de l'iris. La choroïde est une couche très vascularisée alimentant plusieurs structures telles que la rétine et l'iris en nutriments. Comprenant de nombreux pigments de mélanine, elle permet également d'empêcher la diffusion des rayons lumineux à l'intérieur de l'œil. A l'avant de la choroïde se trouve l'iris, une membrane contractile ayant la capacité de modifier le diamètre de l'ouverture en son centre, appelée pupille, pour réguler la quantité de lumière pénétrant dans l'œil en fonction de de la luminosité ambiante, à la manière du diaphragme d'un appareil photo. Les corps ciliaires enfin, ont pour principales fonctions de secréter l'humeur acqueuse, de maintenir le cristallin, qui est attaché à ceux-ci par des ligaments suspenseurs (les zonules de Zinn), ainsi que de modifier la courbure de celui-ci par l'action des muscles ciliaires. Ce phénomène participe à l'accommodation, un mécanisme réflexe permettant de voir nettement à différentes distances en fonction de la déformation du cristallin, une lentille biconvexe souple pouvant se bomber ou s'étirer pour modifier la vergence de l'œil, tel qu'illustré dans la Figure 3.

La dernière tunique, la plus interne, est constituée par la rétine, la couche sensible de l'œil permettant la phototransduction : la transformation de l'information lumineuse en signal électrique, par une cascade d'évènements biochimiques dans les photorécepteurs à l'arrivée de photons.

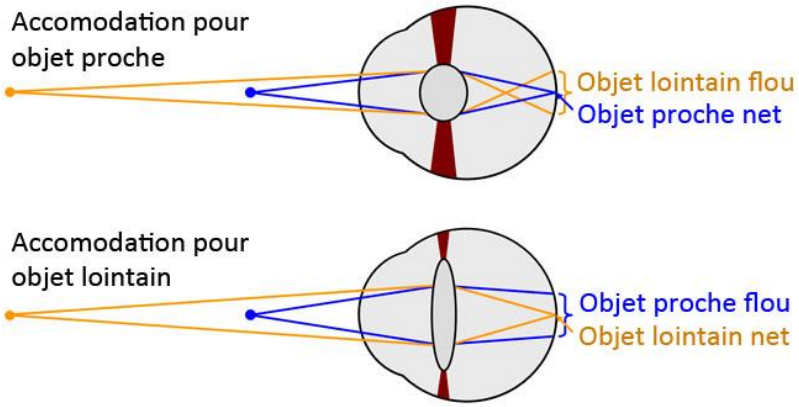


Figure 3 Accommodation de l'œil par l'action de muscles antagonistes dans les corps ciliaires modifiant la convexité du cristallin

1.2 La rétine

La rétine est elle-même composée de différentes couches. La plus externe, après l'épithélium pigmenté, est tapissée de photorécepteurs : les cônes et les bâtonnets. Ces deux types de cellules ont des structures très proches (voir Figure 4), et différent principalement par la taille de leur segment externe. Celui-ci consiste en un empilement de disques membraneux qui renferment des photo-pigments. Lorsque la lumière frappe ces molécules, l'absorption de l'énergie des photons provoque un changement de leur configuration, et une série de réactions chimiques aboutissant à la fermeture de canaux

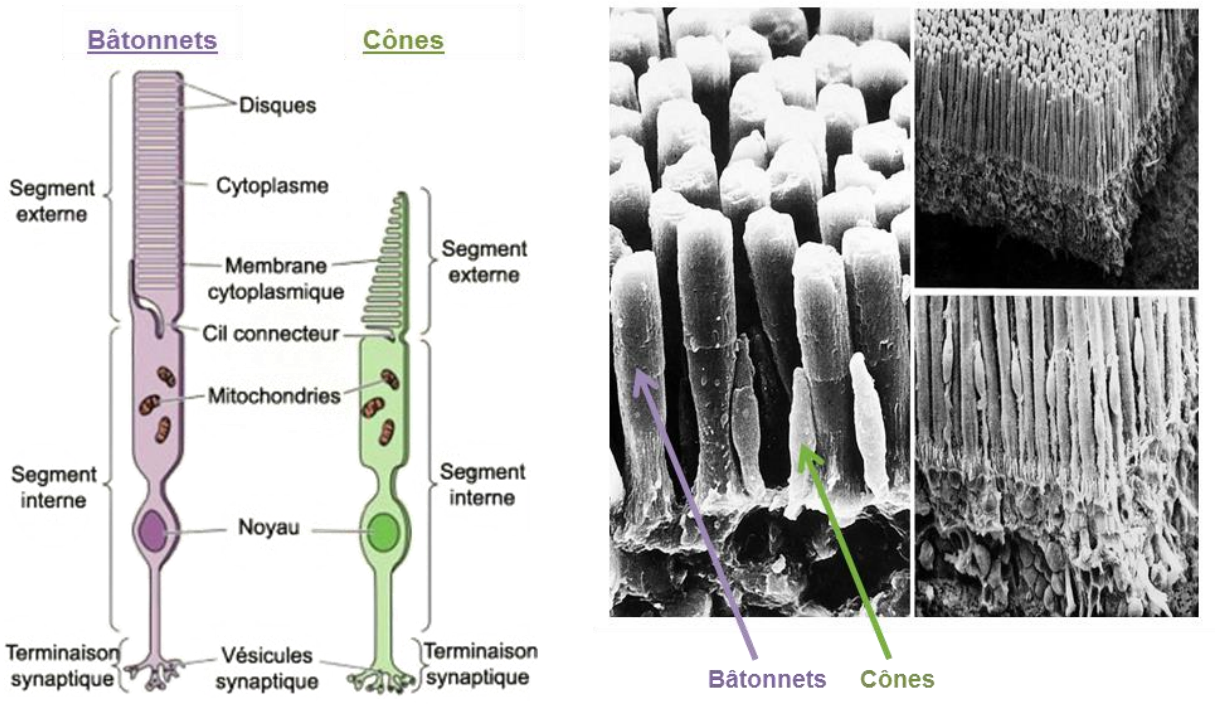


Figure 4 Photorécepteurs: schéma de la structure (à gauche) et photographie microscopique (à droite) des cônes et bâtonnets composant la couche externe de la rétine

sodiques au niveau de la membrane de la cellule. Cette modification de perméabilité va modifier le potentiel de membrane du photorécepteur et permettre l'émission d'un signal nerveux aux cellules de la couche suivante (l'hyperpolarisation de la cellule en fonction de la stimulation lumineuse est illustrée dans la Figure 5).

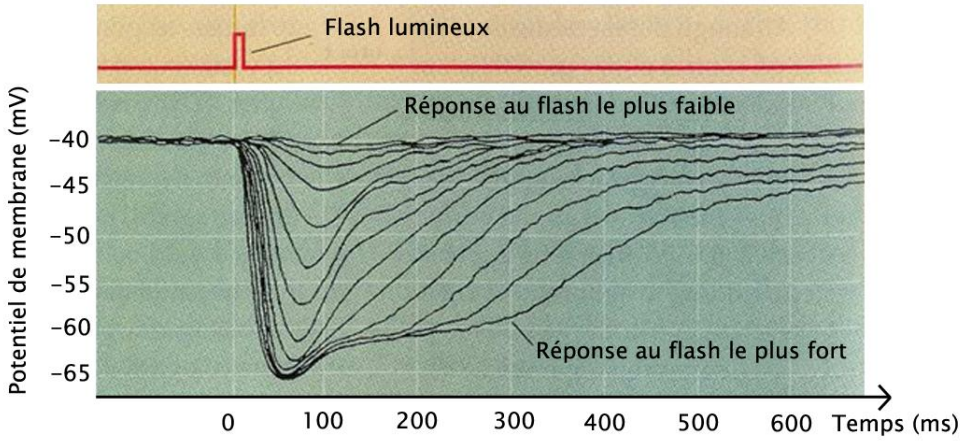


Figure 5 Phototransduction : hyperpolarisation d'un photorécepteur en réponse à une stimulation visuelle

Le nombre plus important de disques dans les bâtonnets induit une sensibilité de ces derniers plus de mille fois supérieure à celle des cônes. Pour cette raison ils permettent la vision à de faibles niveaux d'intensité lumineuse, et participent donc majoritairement au système scopique (du grec, *skotos*, obscurité), saturant lorsque la luminosité est trop forte [Stockman and Sharpe, 2006]. A l'inverse, les cônes nécessitent beaucoup de lumière et interviennent essentiellement en vision diurne (ou photopique). La répartition de ces deux types de photorécepteurs varie grandement au niveau de la rétine. Ainsi, les bâtonnets, au nombre total d'environ 100 millions, sont largement majoritaires dans la périphérie, mais absent de la fovéa ou se concentrent les cônes (estimés à seulement 5 millions à travers la rétine) [Curcio et al., 1990]. Globalement, le nombre total de photorécepteurs décroît plus on s'éloigne de la fovéa, ce qui explique en partie la baisse de l'acuité visuelle avec l'excentricité, illustrée dans la Figure 6.

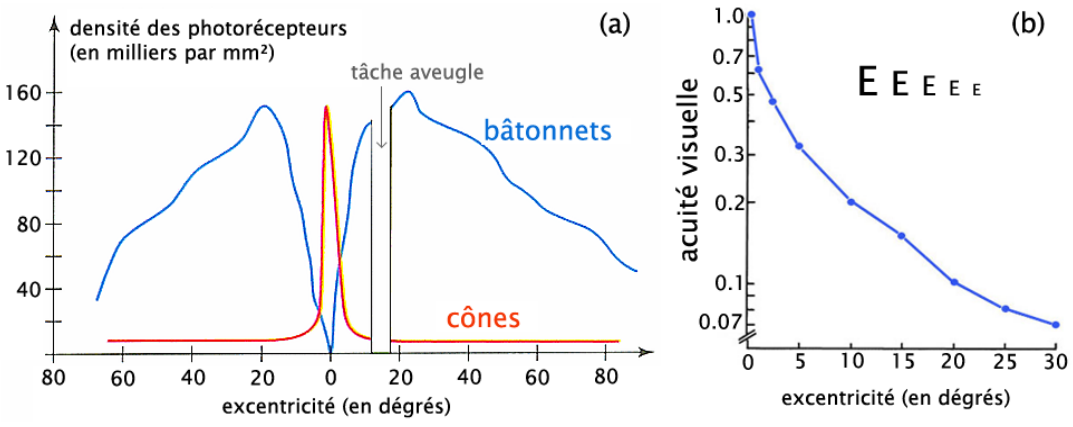


Figure 6 Répartition des photorécepteurs (a) et acuité visuelle (b) selon l'excentricité

Une dernière différence entre cônes et bâtonnets tient au type de photopigments qu'ils contiennent. Chez l'homme il en existe 4 types : la rhodopsine, présente dans les bâtonnets, et trois variétés d'opsine, dans les cônes, ayant chacune des courbes d'absorption de la lumière différentes [Kawamura and Tachibanaki, 2008]. Les bâtonnets ont donc une réponse achromatique, alors que les cônes permettent la perception des couleurs grâce à leurs différentes sensibilités spectrales. On distingue trois catégories de cônes :

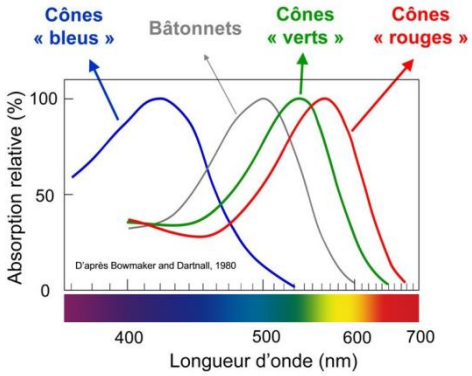


Figure 7 Spectres d'absorption des photorecepteurs

- Les cônes « Bleus » (ou « S, » pour *short wavelengh*), ayant une réponse maximale autour de 440nm ;
- Les cônes « Verts » (ou « M, » pour *middle wavelengh*), ayant une réponse maximale autour de 530nm ;
- Les cônes « Rouges » (ou « L, » pour *long wavelengh*), ayant une réponse maximale autour de 560nm.

Si chacun répond préférentiellement pour une des trois couleurs primaires, leur spectre de sensibilité couvre néanmoins une plage assez large. Pour une couleur donnée les trois types de cônes seront donc simultanément stimulés, mais à des intensités différentes. C'est sur cette combinaison de réponse que repose le modèle trichromatique de la vision humaine [Dacey, 1996; Solomon and Lennie, 2007].

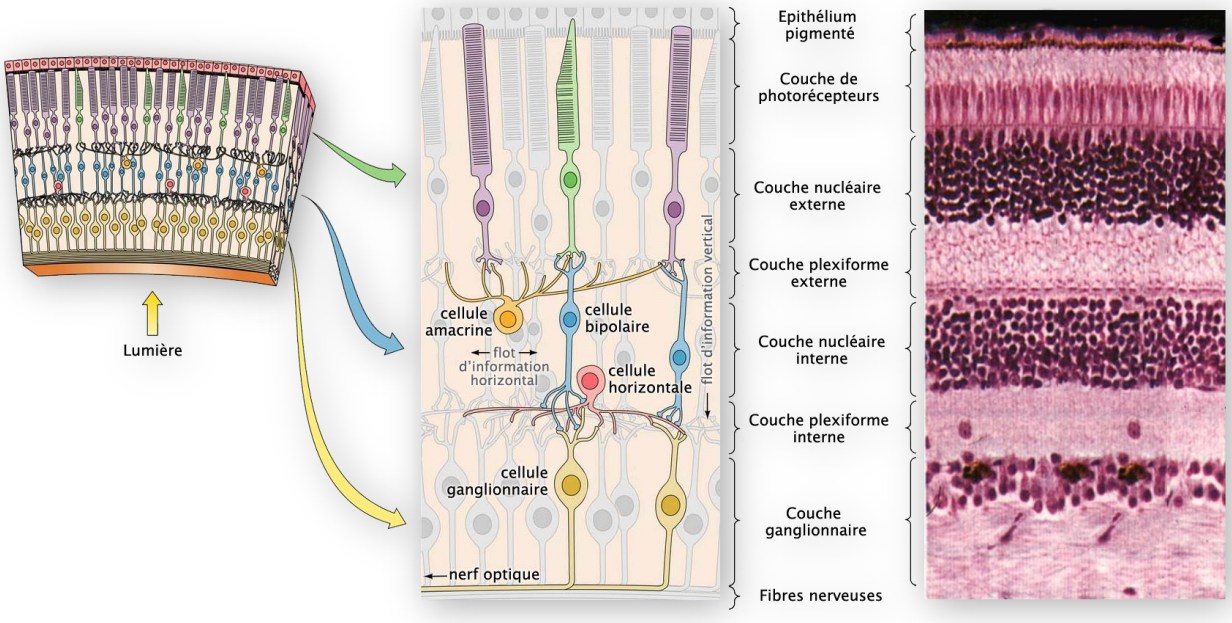


Figure 8 Coupe d'une rétine humaine

A la suite de l'activation des photorécepteurs, le signal nerveux résultant de la phototransduction est propagé aux autres neurones rétiniens. Notons que chez l'homme, les photorécepteurs se trouvent dans la partie externe de la rétine, l'information nerveuse chemine donc dans le sens inverse de la lumière, qui doit d'abord traverser les différentes couches de la rétine pour atteindre les cellules photosensibles. Ces différentes couches, détaillées dans la Figure 8, contiennent en tout 5 catégories de neurones : les photorécepteurs, les neurones bipolaires, ganglionnaires, ainsi que deux types d'interneurones nommés cellules horizontales et amacrines. On dénombre, comme nous l'avons vu, plus de 125 millions de photorécepteurs, alors qu'après le dernier relais synaptique rétinien, on ne compte qu'entre 1 et 2 millions de neurones ganglionnaires, dont les axones composent le nerf optique, transmettant l'information visuelle au cerveau. L'architecture et les traitements effectués dans la rétine vont permettre une compression de l'information visuelle en codant les motifs visuels sous forme de centre/pourtour afin de de diminuer la quantité de signal à transmettre jusqu'au cortex visuel [Wandell, 1995].

Cette transformation dans la nature de l'information véhiculée par les cellules de la rétine est le résultat de sommations de leurs champs récepteurs. On nomme champ récepteur (CR) d'un neurone la région de l'espace pour laquelle la présentation d'un stimulus adapté induit une modification de son activité [Hubel, 1994]. Les plus petits champs récepteurs du système visuel correspondent aux photorécepteurs, dont l'activité ne dépend que d'une petite portion du champ visuel. Grâce aux interneurones horizontaux, les réponses de nombreux photorécepteurs vont converger vers des cellules bipolaires, selon une organisation spatiale particulière, en forme d'anneaux concentriques [Dacey et al., 2000; Piccolino, 1995], illustrée dans la Figure 9.

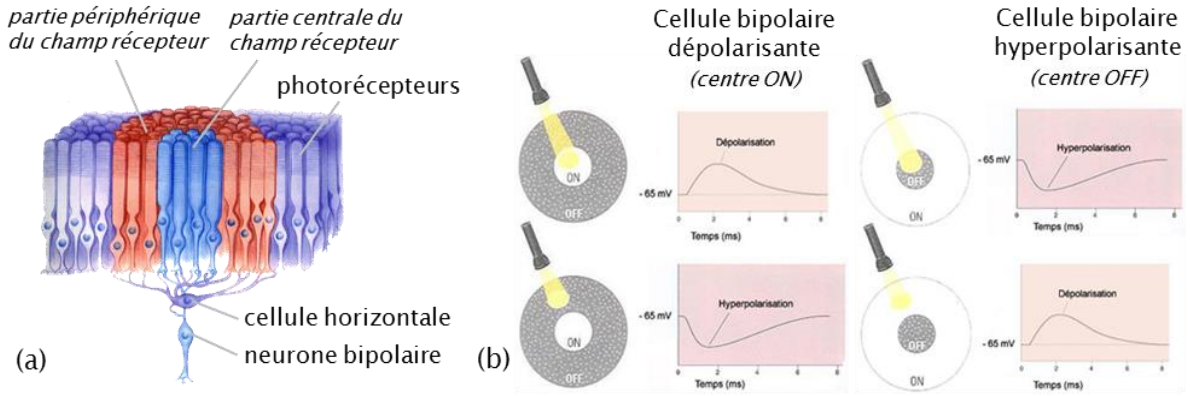


Figure 9 Organisation (a) et réponse (b) du champ récepteur de neurones bipolaires

Chaque neurone bipolaire reçoit des connexions synaptiques directes de photorécepteurs (de un au centre de la fovéa jusqu'à plusieurs milliers dans la périphérie de la rétine), mais également des afférences de cellules horizontales, qui sont elles-mêmes

reliées à un ensemble de photorécepteurs entourant le groupe central [Buser and Imbert, 1987]. Ces cellules horizontales ont une action inhibitrice et permettent un antagonisme fonctionnel entre le centre et le pourtour, on parle donc d'inhibition latérale. Le CR des neurones bipolaire correspond à la somme des CRs individuels des cônes ou bâtonnets avec lesquels ils communiquent et s'organise en deux parties :

- un champ récepteur central constitué de l'information qui transite directement des photorécepteurs aux cellules bipolaires,
- ainsi qu'un champ récepteur périphérique qui reçoit l'information passant par les cellules horizontales.

Il est aussi important de noter qu'on trouve deux types de neurones bipolaires, aux réponses opposées. L'un, dépolarisant, est dit « Centre ON » car il s'active lorsque la partie centrale de son CR est stimulée, et le deuxième, hyperpolarisant (ou « Centre Off »), répond à l'éclairement de sa partie périphérique, tel que décrit dans la Figure 9. Ces neurones, toujours associés par paire, fonctionnent de manière indépendante et parallèle. Leurs différences s'expliquent par la nature des synapses qui les lient aux photorécepteurs, pouvant être à rubans ou superficielles, entraînant dans un cas l'inhibition, dans l'autre l'excitation du neurone lors de la libération de glutamate par les cônes ou bâtonnets.

Le dernier étage de la chaîne de traitement de la rétine est constitué par les neurones ganglionnaires. Ceux-ci possèdent des champs récepteurs similaires aux neurones bipolaires, de type centre-pourtour, et s'organisent également en paires antagonistes ON et OFF[1]. Par le biais des cellules amacrines, et des synapses directes avec la couche précédente, de nombreux neurones bipolaires vont converger vers une même cellule ganglionnaire, augmentant une nouvelle fois la taille des champs récepteurs [Masland, 2001]. Ce nombre de connexions dépend en partie de l'excentricité, la convergence étant plus forte dans la périphérie de la rétine, mais aussi du type de neurones ganglionnaires [Rodieck, 1998]. Parmi la vingtaine de cellules ganglionnaires identifiées chez l'homme on distingue en effet trois Grandes catégories, l'origine des voies de traitements parallèles magno, parvo et koniocelléllulaire :

- Les neurones parasols, ou alpha, dits de type M (pour magnocellulaire), se caractérisent par une large arborisation dendritique et de grands champs récepteurs. Ils sont présents dans toute la rétine et nécessitent des stimulations de niveaux lumineux photopiques. Recevant les informations de cônes rouges et verts sans en faire la distinction, leur réponse est achromatique. On estime qu'ils représentent environ 10 % des fibres du nerf optique [Lee, 1996].

[1] Les cellules bipolaires ON projettent vers des neurones ganglionnaires ON, les OFF vers des OFF.

- Les neurones nains de type P[1] (parvocellulaire), aux champs récepteurs plus réduits, sont largement majoritaires en nombre (plus de 80%) [Wässle, 2004]. Ils reçoivent séparément les afférences de cônes rouges et verts (par le biais de neurones bipolaires de cônes nains), et permettent le codage de l'antagonisme rouge-vert, premier canal à l'origine de la vision chromatique. Par l'intermédiaire de cellules amacrines, les cellules ganglionnaires de type P véhiculent également l'information provenant de cellules bipolaires de bâtonnets, participant ainsi au système scotopique lorsque la luminosité est trop faible pour activer les cônes.

- Les neurones bistratifiés de type K (koniocellulaire), ou cellules gamma, ont été découverts beaucoup plus tard que les deux précédents, du fait de leur corps cellulaires extrêmement petits [P. R. Martin et al., 1997]. Ils représentent environ 5 % de l'ensemble des cellules ganglionnaires et sont responsables de l'antagonisme bleu-jaune, deuxième canal permettant la vision des couleurs [Dacey and Lee, 1994]. Ils communiquent d'une part avec des cellules bipolaires de cônes bleus (très minoritaires [Calkins, 2001]), et d'autres part avec des cellules bipolaires cônes L et M diffuses.

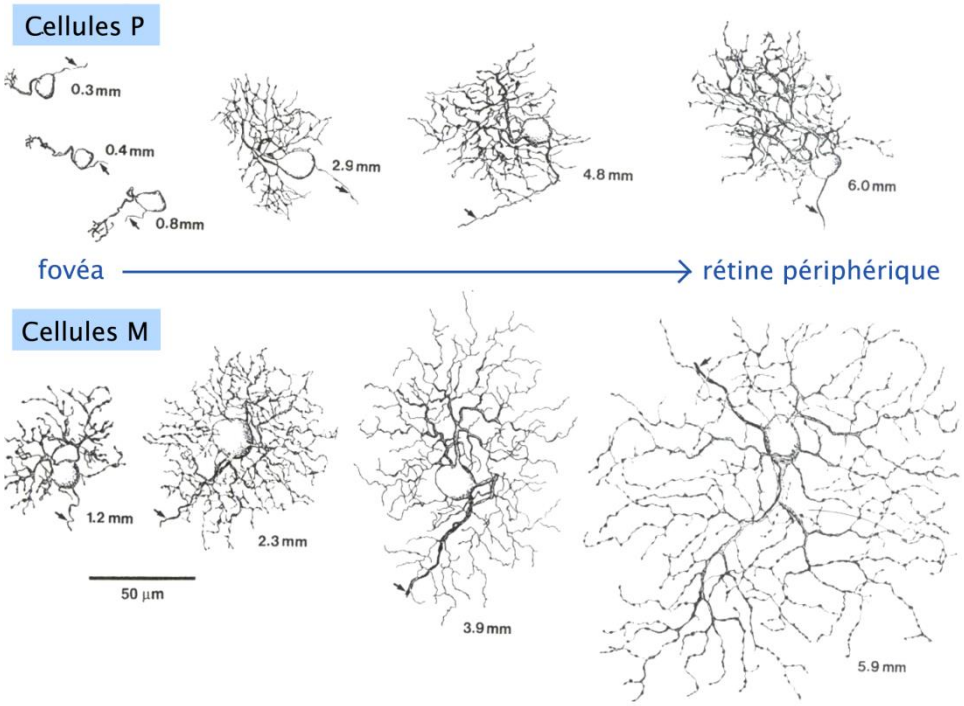


Figure 10 Taille moyennes des cellules ganglionnaires de type P et M selon l'excentricité

[1] Aussi appelés cellules bêta.

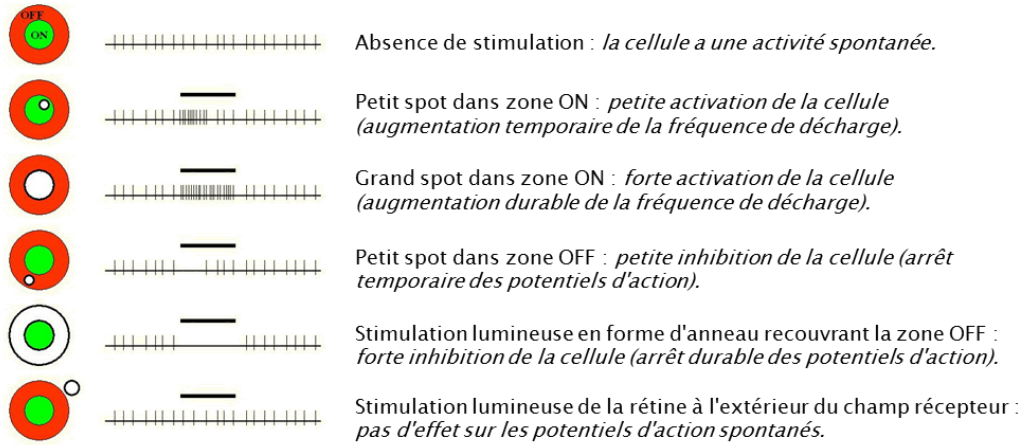


Figure 11 Activité d'un neurone ganglionnaire centre ON en fonction de différentes stimulations lumineuses

Les axones de ces différentes cellules ganglionnaires se réunissent au niveau de la papille[1], où ils traversent la rétine pour former le nerf optique, transmettant l'information visuelle au cerveau. Pour que ces signaux soient transmis sans perte d'information, des potentiels d'action (PA) doivent être générés et propagés le long de ces fibres optiques. Les PA sont des inversions brutales et transitoires du potentiel de membrane, qui se propagent le long des axones sans atténuation. Ils obéissent à la loi du tout au rien : si le seuil de dépolarisation n'est pas atteint, aucun PA n'est émis, en revanche dès qu'il est dépassé, la réponse est immédiatement maximale. C'est par conséquent leur fréquence et/ou leur latence qui permettent de coder l'intensité de l'activité du neurone [Kuffler, 1953], tel qu'illustré dans la Figure 11. Les neurones ganglionnaires sont les seules cellules de la rétine à émettre des PA, les autres ne contenant pas de canaux Na+ sensibles au voltage, et ne transmettant l'information que par une simple déploration ou hyperpolarisation (voir Figure 12).

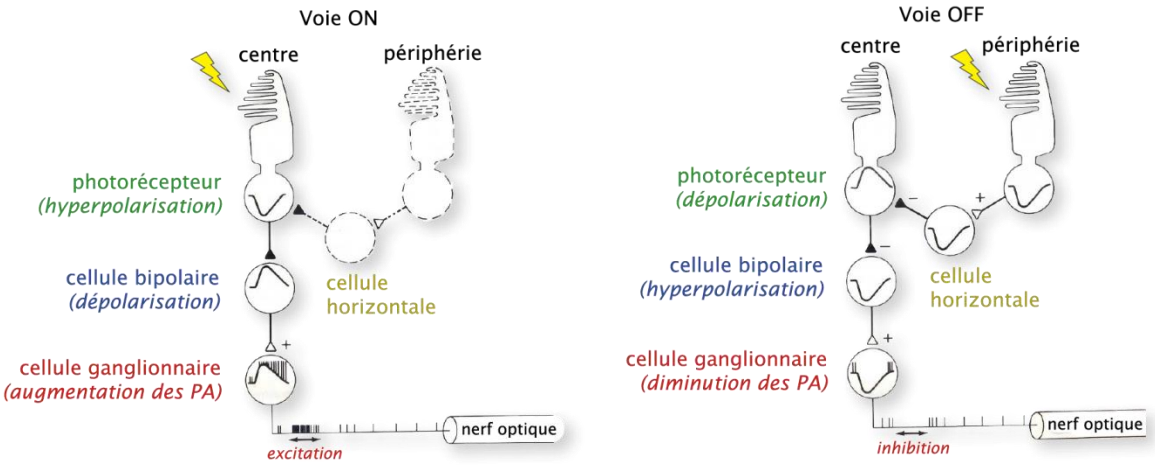


Figure 12 Réponses des différents étages de ma rétine à une stimulation lumineuse

[1] Aussi appelée tâche aveugle, du fait de l'absence de photorécepteurs dans cette région de la rétine.

1.3 Voies visuelles

Les traitements effectués par la rétine permettent, comme nous venons de le voir, de compresser l'information visuelle par la convergence des champs récepteurs, augmentant à chaque étape étage rétinien, et par l'extraction de contrastes locaux, grâce aux inhibitions latérales, entre des zones de luminosité ou de chrominance différentes. Ces mécanismes permettent de recoder l'information visuelle de façon optimale, sous une forme décorrélées [Atick and Redlich, 1992; Dan et al., 1996], les représentations au niveau des photorécepteurs souffrant en effet d'une forte redondance, dues à de hautes corrélations spatiales et temporelle de leurs activations.

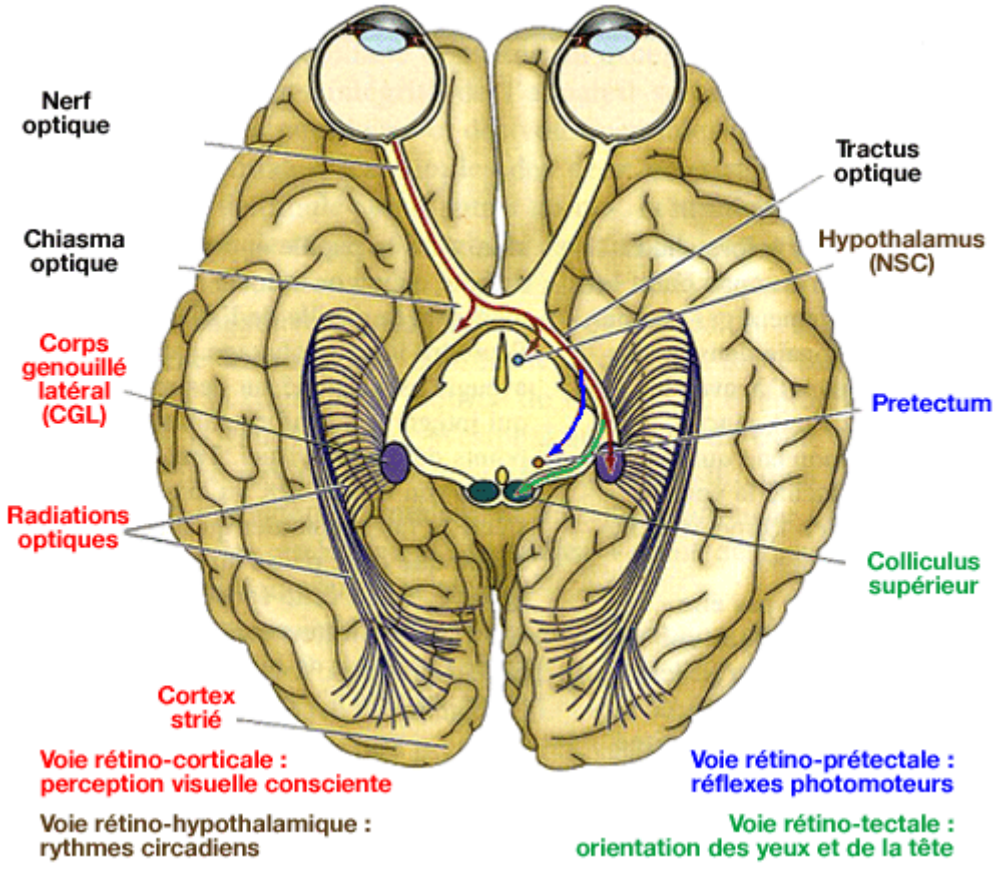


Figure 13 Structures impliquées dans les principales voies visuelles

Il est également important de souligner la surreprésentation de la vision centrale (macula et fovéa) dans ce signal visuel, où la densité de photorécepteurs est maximale, et où les circuits sont faiblement convergents (seulement un ou quelques photorécepteurs par neurone ganglionnaire, pour plusieurs milliers en périphérie) [Sterling, 2004]. Ces

mécanismes permettent un traitement bien plus fin autour du point de fixation grâce à l'acuité maximale de la fovéa et des propriétés fonctionnelles différentes de celles de du reste du champ visuel. Les informations véhiculées par les nerfs optiques vont ensuite se propager dans le cerveau en empruntant plusieurs voies visuelles parallèles, tel qu'illustré dans la Figure 13.

1.3.1 Voies retino-tectale, retino-prétectale, et optique accessoire

Environ 10 % des fibres des axones ganglionnaires vont par exemple se projeter dans le colliculus supérieur (CS) et le prétectum (on parle respectivement des voies retino-tectales et retino-prétectales). Ces structures du mésencéphale constituent les centres visuels chez de nombreuses espèces telles que les poissons, amphibiens ou reptiles. Chez les mammifères en revanche, le néocortex est devenu la région principale de traitement pour la majorité des fonctions visuelles. Les noyaux prétectaux et le CS conservent néanmoins un rôle majeur dans l'intégration sensorimotrice, recevant à la fois des informations visuelles, auditives, réticulaires, vestibulaires et cérébelleuses, organisées sous formes de cartes visuotopiques. Ils sont en particulier impliqués dans le contrôle des mouvements oculaires, comme la programmation de saccades et différents mécanismes réflexes listés ci-dessous:

- Réflexe optocinétique : stabilise le regard sur une cible en mouvement pour une vision nette de celui-ci.
- Mouvements vestibulo-oculaires : compensent les rotations rapides de la tête par des saccades de directions opposées. Notons qu'en plus de la voie retino-prétactale, une seconde, appelée voie optique accessoire, participe aussi au maintien de la direction du regard lors des mouvements de la tête. Celle-ci est composée de petits groupements cellulaires (les noyaux terminaux dorsaux, latéraux, et médians), recevant des afférences de neurones ganglionnaires sélectifs à des directions spécifiques, qui permettent de corriger, si nécessaire, les erreurs de mouvements vestibulo-oculaires.
- Réflexe pupillaire (ou photomoteur) : les projections des aires prétectacles sur les voies parasympathiques (oridoconstrictrices) et orthosympathiques (iridodilatractices) modifient de façon automatique l'ouverture de la pupille en fonction des conditions de luminosité.
- Réflexe d'accommodation : par l'envoi de commandes motrices aux muscles ciliaires de l'œil, le noyau d'Edinger-Westphal permet de modifier la courbure du cristallin en fonction de la distance de l'objet fixé afin d'en assurer sa mise à point.

- Mouvements de vergence : les axes de nos deux yeux peuvent être presque parallèles, lorsque nous regardons des objets lointains, ou au contraire converger fortement en fixant des objets proches. Par des mécanismes similaires aux refléxes d'accommodation, le noyau d'Edinger-Westphal participe également à ces mouvements de vergence médiés par les muscles droits médiaux.

1.3.2 Voie tecto-pulvinarienne

Le pulvinar est un noyau localisé dans la région postérieure du thalamus principalement impliqué dans la vision [Robinson and Cowie, 1997]. Il reçoit des afférences d'autres structures sous-corticales voisines (telles que le CS, le prétectum, ou le corps géniculé latéral) mais aussi des connexions directes avec la rétine. Son rôle dans l'analyse des informations visuelles a longtemps été méconnu et reste encore relativement flou [Grieve et al., 2000]. Néanmoins, un nombre croissant d'études semblent contredire les modèles traditionnels qui le considéraient comme un simple relais passif de l'information cheminant vers le cortex.

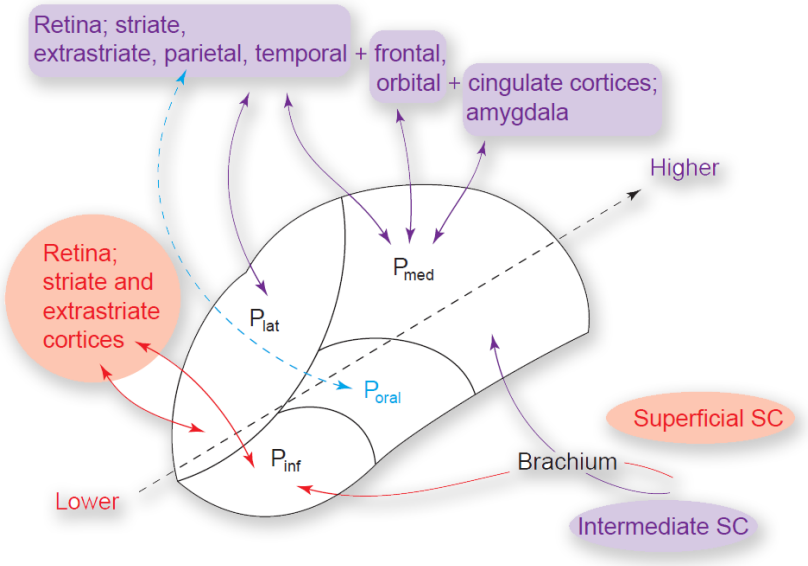


Figure 14 Connectivité du pulvinar chez le primate (tirée de [Grieve et al., 2000])

De multiples connexions bidirectionnelles ont en effet été identifiées entre le pulvinar et la plupart des aires visuelles, aussi bien dans le cortex strié, effectuant des traitements précoces « bas niveau », que dans des structures supérieures des régions temporales ou pariétales [Berman and Wurtz, 2008; Leh et al., 2007], comme illustré dans la Figure 14. Différentes études en IRM fonctionnelle ont montré que certains neurones du pulvinar

pouvaient éliciter des réponses sophistiquées que l'on pensait limitées au cortex visuel. [Vuilleumier et al., 2003]. Il semble d'ailleurs que leurs champs récepteurs possèdent les mêmes propriétés que ceux des cellules corticales auxquelles ils sont connectés [Rigoulot, 2008].

Au vu des connaissances actuelles, le pulvinar apparaît donc être impliqué dans les mécanismes oculomoteurs [Chalupa, 1977], attentionnels [Saalmann et al., 2012; Yates, 2012], et dans certains traitements visuels inconscients [Mulckhuyse and Theeuwes, 2010; Vuilleumier et al., 2003]. De récents enregistrements intracrâniens ont aussi montré un nombre conséquent de neurones dans la partie médiane et dorsolatérale du pulvinar présentant des réponses sélectives à des images de visages, de mains ou de serpents [Le et al., 2013]. Leurs latences très faibles et leur sensibilité aux basses fréquences spatiales suggèrent l'existence de circuits sous-corticaux impliqués dans la détection rapide de cibles écologiquement importantes, organisées en une architecture de nature feedfoward[1] [Johnson, 2005; Sewards and Sewards, 2002; Vuilleumier et al., 2003; Ward et al., 2005].

1.3.3 Voie rétino-hypothalamique

Une petite partie des cellules ganglionnaires de la rétine se projettent également dans les noyaux supra-chiasmatiques de l'hypothalamus, assurant avec l'épiphyse la synchronisation des rythmes circadiens de l'organisme en fonction de l'alternance jour-nuit. Ils influent notamment sur de nombreux aspects physiologiques tels que les variations de la température, les sécrétions hormonales ou les états de veille et de sommeil.

1.3.4 Voie geniculo-striée

Avec près de 90 % des fibres optiques se projetant dans les corps géniculés latéraux, la voie geniculo-striée[2] constitue chez le primate la voie visuelle primaire. Avant d'atteindre le thalamus, les axones des cellules ganglionnaires constituant les nerfs optiques partant de chaque œil se réunissent au niveau du chiasma optique, une zone de décussation permettant de regrouper les informations relatives à chaque hémichamp visuel, et de les acheminer dans l'hémisphère opposé par les bandelettes optique (ou tractus). Les faisceaux maculaires vont donc se diviser à ce niveau, les fibres temporales de chaque rétine poursuivant un trajet direct (ou ipsilatéral), tandis que les fibres nasales suivent une voie controlatérale, tel qu'illustré dans la Figure 15. Ces tractus se terminent dans les corps géniculés latéraux

[1] On parle aussi de traitements bottum-up, ou ascendants.
[2] Egalement appelée voie retino-geniculo-corticale.

(CGL), des relais thalamiques constitués de 6 couches concentriques. Chacune de ces couches ne reçoit des afférences que d'un seul œil, les fibres optiques issues de l'œil ipsilatéral terminant dans les couches 2, 3 et 5, dans que celle de l'œil controlatéral se projettent dans les couches 1, 4 et 6 (voir la Figure 16 décrivant l'organisation laminaire du CGL).

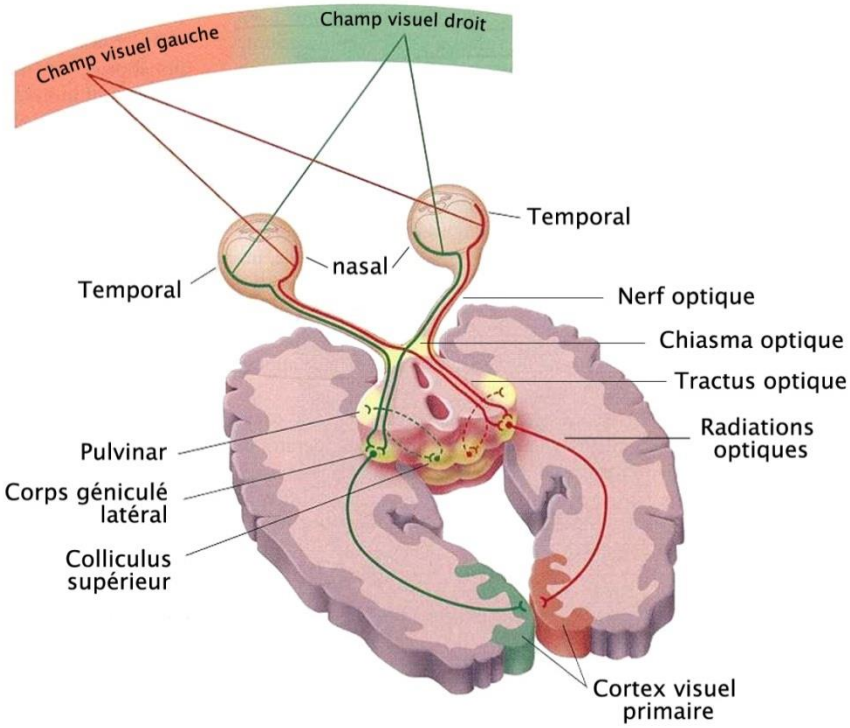


Figure 15 Voie géniculo-striée (informations passant de la rétine au corps géniculé latéral, puis au cortex visuel primaire)

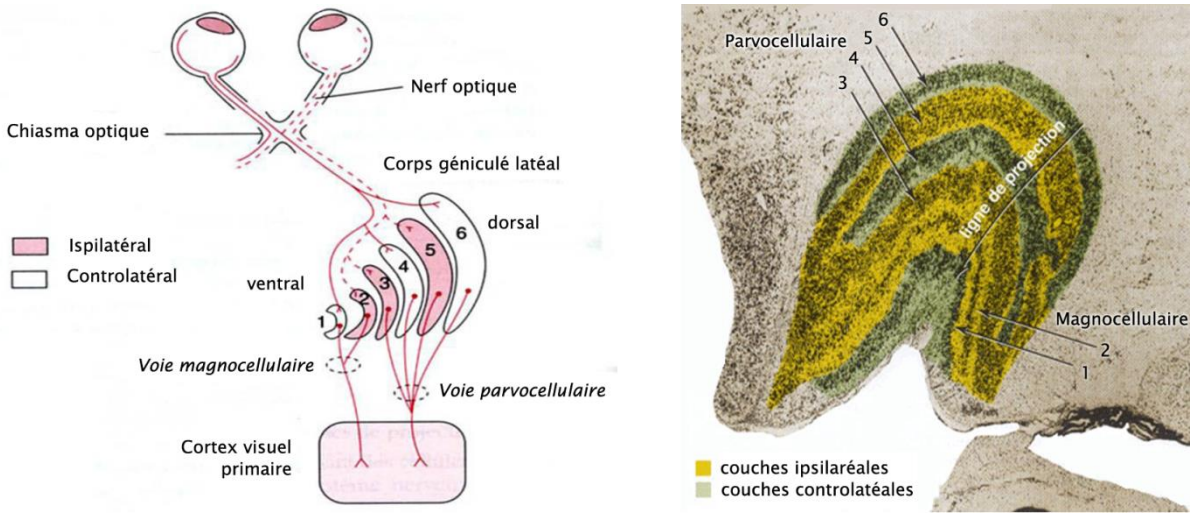


Figure 16 Organisation et projections du corps géniculé latéral

En plus de la séparation binoculaire, on observe également une ségrégation de l'information visuelle liée aux types de neurones ganglionnaires afférents [Bullier, 2002]. Les cellules M, localisées dans la partie ventrale du CGL (couches 1 et 2) reçoivent ainsi les projections des neurones ganglionnaires de type M (ou cellules alpha), tandis que les cellules P des couches dorsales (3, 4, 5 et 6), reçoivent celles des neurones ganglionnaires de type P (ou cellules bêta). Les cellules K enfin, forment les sous-couches koniocellulaires intercalées entre les couches parvo et magnocellulaires [Hendry and Reid, 2000].

Du point de vue fonctionnel, le CGL est traditionnellement considéré comme un simple relais thalamique de l'information rétinienne, transmettant celle-ci par le biais des radiations optiques à l'aire visuelle primaire (ou V1) pour un traitement plus poussé. Une grande majorité (près de 80 %) des cellules du CGL semblent n'être effectivement que de simples liaisons, recevant leurs entrées des cellules ganglionnaires, et projetant un axone dans le cortex strié, sans développer de nouvelles propriétés de sélectivité [Bullier, 2002]. On suppose néanmoins que ces cellules relais puissent être soumises à certaines modulations par l'action d'interneurones, de rétroprojections corticales, ou d'afférences d'autres structures sous-corticales, sans pour autant que ces mécanismes soient réellement connus [Guillery and Sherman, 2002].

La plupart des champs récepteurs observés dans le CGL sont, en conclusion, très similaires à ceux des cellules ganglionnaires, répondant aux contrastes de luminosité ou de teinte illustrés dans la Figure 17. Par la différenciation des couches parvo et magnocellulaires, l'information transmise au cortex visuel emprunte deux voies distinctes et parallèles[1], aux propriétés différentes :

- La voie magnocellulaire, plus sensible aux faibles contrastes et aux basses fréquences spatiales, possède de grands champs récepteurs, achromatiques, dont les réponses sont transitoires et rapides (10 à 20 ms plus tôt que les cellules P).

- La voie parvocellulaire, qui prédomine en vision centrale, comprend de petits champs récepteurs, permettant une analyse fine des détails par leur sensibilité aux hautes fréquences spatiales et aux différences chromatiques. Les cellules P présentent une réponse tonique (maintenue durant la stimulation) mais une conductivité plus lente. Elles sont aussi moins sensibles aux faibles contrastes et aux mouvements.

[1] Nous n'aborderons pas la voie koniocellulaire, dont les propriétés restent assez méconnues, notamment du fait de la difficulté à enregistrer des cellules si petites.

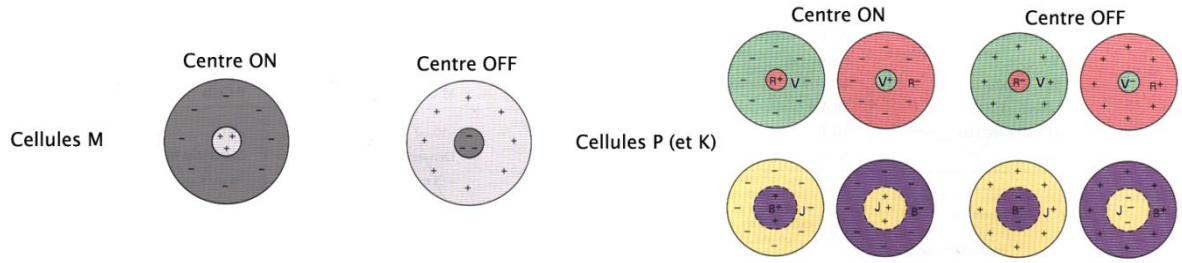


Figure 17 Champs récepteurs des cellules M et P du corps géniculé latéral

Les radiations optiques constituent la dernière étape de la voie géniculo-striée, acheminant le signal visuel transmis par le CGL au cortex visuel primaire, premier relais cortical des informations de la rétine, où les représentations visuelles se caractérisent par une organisation rétinotopique[1] et des champs récepteurs sensibles aux orientations (voir Figure 18).

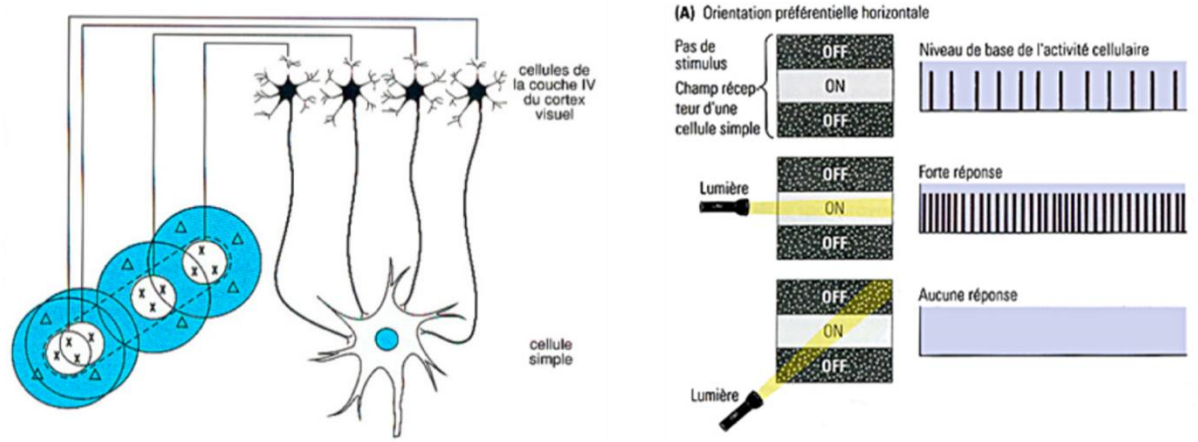


Figure 18 Construction du champ récepteur d'une cellule simple de V1

A partir du cortex strié l'information se propage dans un vaste réseau d'aires visuelles et associatives de la partie postérieure du cerveau (voir Figure 19). Malgré la complexité des interconnexions et des flux d'informations entre ces différentes régions cérébrales, deux grandes voies fonctionnelles se dégagent : la voie ventrale et la voie dorsale, respectivement alimentées par les neurones parvo et magnocellulaires [Livingstone and Hubel, 1988]. Partant toutes deux du cortex visuel primaire (l'aire V1), la première se termine dans le cortex inféro-temporal et la seconde au niveau du cortex pariétal [Noë and Thompson, 2002], comme illustré dans la Figure 20. Si les opinions diffèrent quant aux fonctions exactes de ces deux systèmes l'hypothèse la plus répandue est que la voie ventrale est spécialisée dans la reconnaissance alors que la voie dorsale est principalement impliquée (on parle souvant du « What and Where ») [Mishkin et al., 1983].

[1] La rétinotopie est la propriété d'une aire visuelle qui consiste à représenter une partie de la rétine (et donc du champ visuel) de manière ordonnée sur la surface corticale. Deux neurones proches sur la surface corticale seront ainsi activés par des régions voisines du champ visuel.

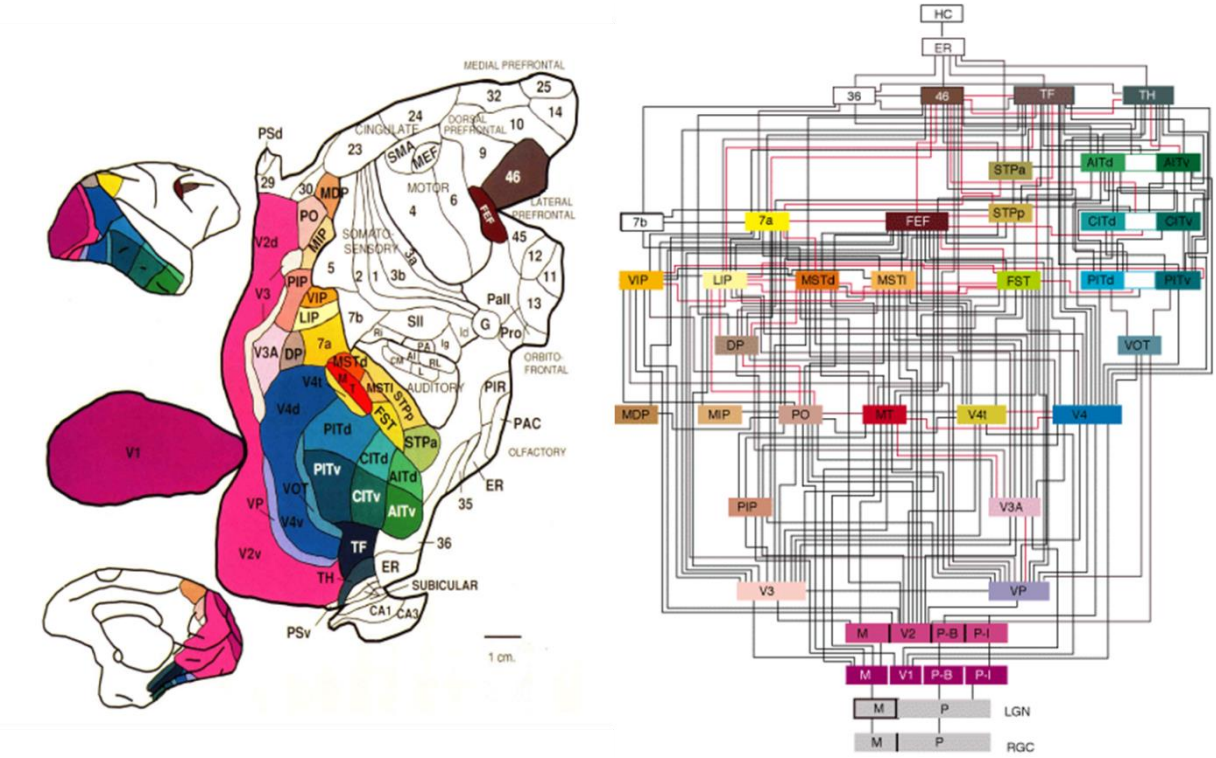


Figure 19 Anatomie et connectivité des aires visuelles

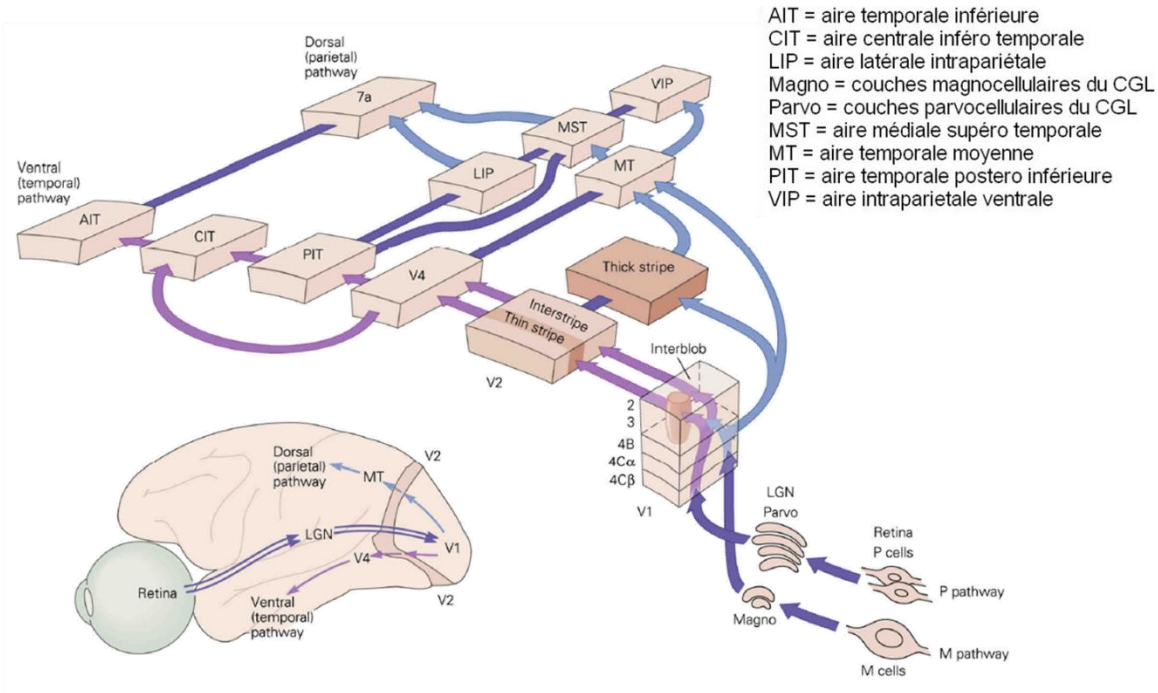


Figure 20 Vue schématiques des deux voies ventrales et dorsales chez le primate

2. Logiciels relatifs à Spikenet MultiRes

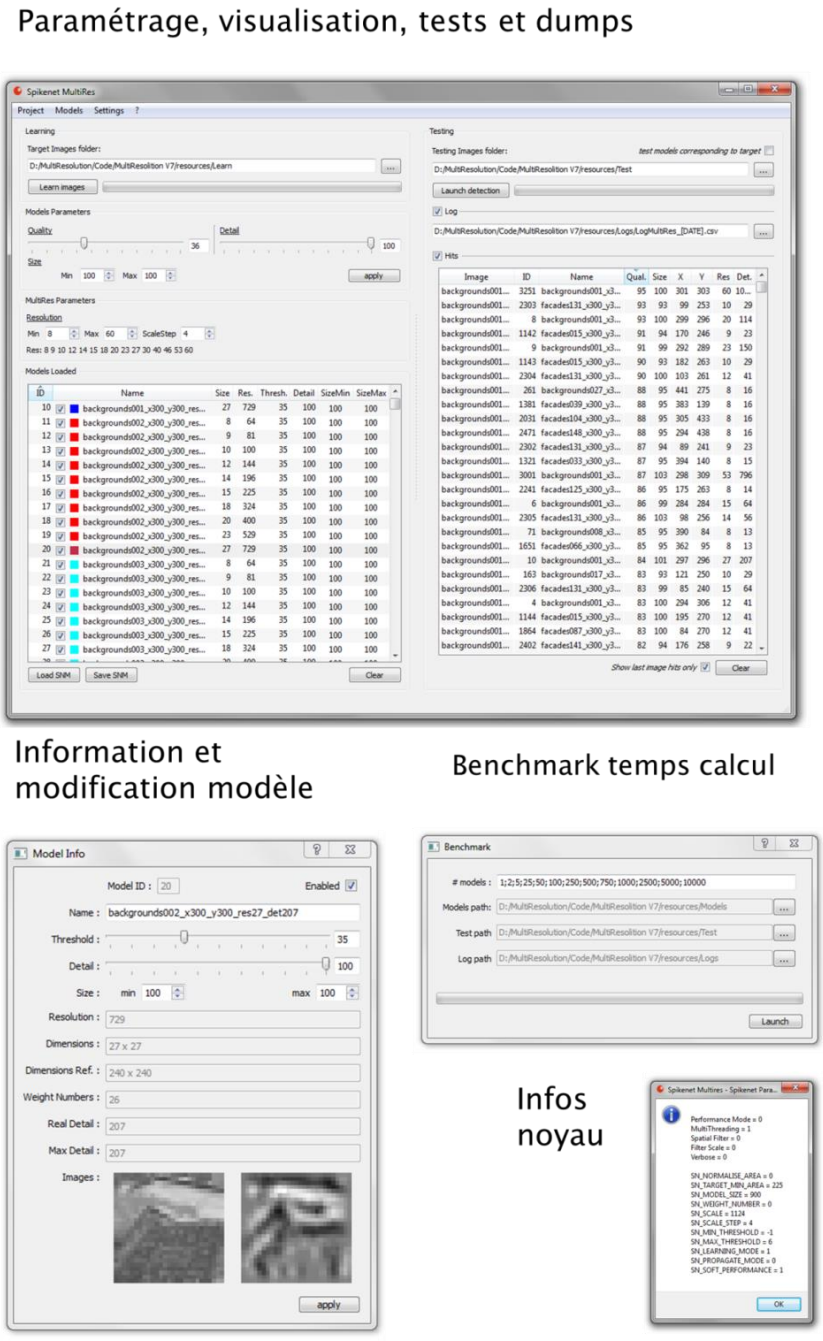


Figure 21 Principaux composants de l'interface du logiciel exploitant le noyau MultiRes

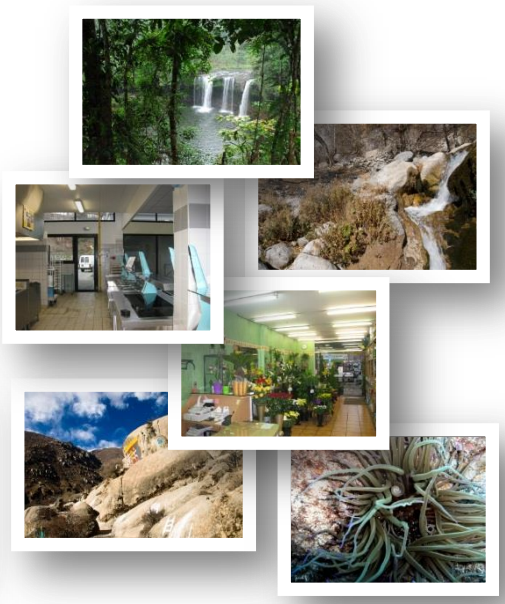
	imName	imPath	modelID	modelName	modelRes	modelDetail	hitQual	hitX	hitY	hitScale
	backgrounds001.jpg	D:/MultiResolution/Code/MultiReso	841	backgrounds061_x300_y300_res08_det16	8	15	93	126	505	105
	backgrounds001.jpg	D:/MultiResolution/Code/MultiReso	841	backgrounds061_x300_y300_res08_det16	8	15	93	142	505	105
	backgrounds001.jpg	D:/MultiResolution/Code/MultiReso	9	backgrounds001_x300_y300_res30_det255	30	255	92	300	300	100
	backgrounds001.jpg	D:/MultiResolution/Code/MultiReso	6	backgrounds001_x300_y300_res20_det114	20	114	92	300	300	100
	backgrounds001.jpg	D:/MultiResolution/Code/MultiReso	3333	facades139_x300_y300_res08_det16	8	12	92	457	394	105
	backgrounds001.jpg	D:/MultiResolution/Code/MultiReso	3333	facades139_x300_y300_res08_det16	8	12	92	473	394	105
	backgrounds001.jpg	D:/MultiResolution/Code/MultiReso	3333	facades139_x300_y300_res08_det16	8	12	92	126	521	105
	backgrounds001.jpg	D:/MultiResolution/Code/MultiReso	1205	backgrounds087_x300_y300_res08_det16	8	12	92	63	173	105
	backgrounds001.jpg	D:/MultiResolution/Code/MultiReso	1205	backgrounds087_x300_y300_res08_det16	8	12	92	78	173	105
	backgrounds001.jpg	D:/MultiResolution/Code/MultiReso	1205	backgrounds087_x300_y300_res08_det16	8	12	92	78	189	105
	backgrounds001.jpg	D:/MultiResolution/Code/MultiReso	1205	backgrounds087_x300_y300_res08_det16	8	12	92	78	205	105
	backgrounds001.jpg	D:/MultiResolution/Code/MultiReso	659	backgrounds048_x300_y300_res08_det16	8	12	92	205	63	105
	backgrounds001.jpg	D:/MultiResolution/Code/MultiReso	659	backgrounds048_x300_y300_res08_det16	8	12	92	299	63	105
	backgrounds001.jpg	D:/MultiResolution/Code/MultiReso	225	backgrounds017_x300_y300_res08_det16	8	13	92	142	173	105
	backgrounds001.jpg	D:/MultiResolution/Code/MultiReso	225	backgrounds017_x300_y300_res08_det16	8	13	92	142	189	105
	backgrounds001.jpg	D:/MultiResolution/Code/MultiReso	225	backgrounds017_x300_y300_res08_det16	8	13	92	126	252	105
	backgrounds001.jpg	D:/MultiResolution/Code/MultiReso	225	backgrounds017_x300_y300_res08_det16	8	13	92	126	331	105
	backgrounds001.jpg	D:/MultiResolution/Code/MultiReso	7	backgrounds001_x300_y300_res23_det150	23	150	91	297	297	102
	backgrounds001.jpg	D:/MultiResolution/Code/MultiReso	2201	facades058_x300_y300_res12_det41	12	41	90	434	72	103
	backgrounds001.jpg	D:/MultiResolution/Code/MultiReso	3334	facades139_x300_y300_res10_det29	10	29	90	480	396	100
	backgrounds001.jpg	D:/MultiResolution/Code/MultiReso	2578	facades085_x300_y300_res10_det29	10	29	90	396	72	100
	backgrounds001.jpg	D:/MultiResolution/Code/MultiReso	2578	facades085_x300_y300_res10_det29	10	29	90	420	72	100
	backgrounds001.jpg	D:/MultiResolution/Code/MultiReso	2578	facades085_x300_y300_res10_det29	10	29	90	444	72	100

Figure 22 Extraits des logs générés pour l'évaluation du noyau MutltiRes

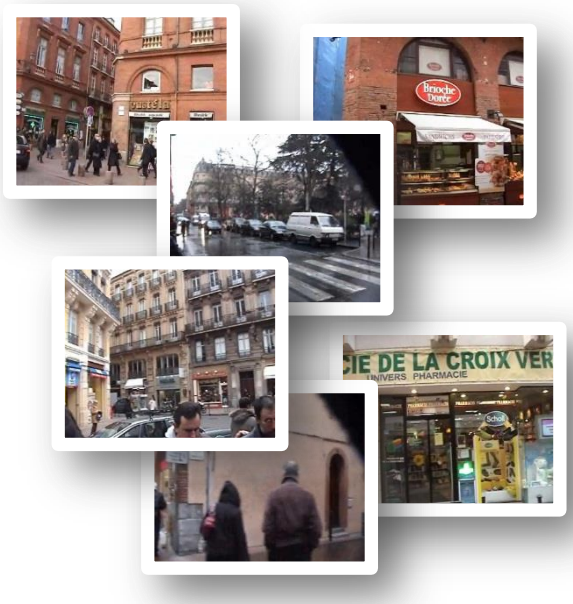
3. Images des benchmarks MultiRes

150 images “facades”

100 images “background”
(images complexes)



250 images “navig”
(images prises en centre-ville, potentiellement proches des images “facades”)



350 images “Films”
(prises aléatoirement dans des films)

4. Listes des publications

G. Parseihian, A. Brilhault, et F. Dramas, « NAVIG: An object localization system for the blind », présenté à 8th International Conference on Pervasive Computing, Helsinki, Finland, 2010.

S. J. Thorpe, A. Brilhault, et J.-A. Perez-Carrasco, « Suggestions for a biologically inspired spiking retina using order-based coding », présenté à Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on, Paris, France, 2010, p. 265-268.

A. Brilhault, S. Kammoun, O. Gutierrez, P. Truillet, et C. Jouffrais, « Fusion of Artificial Vision and GPS to Improve Blind Pedestrian Positioning », présenté à 4th IFIP International Conference on New Technologies, Mobility and Security (NTMS), Paris, France, 2011, p. 1 -5.

A. Brilhault, M. Mathey, N. Jolmes, et S. J. Thorpe, « Measuring the receptive field sizes of the mechanisms underlying ultra-rapid saccades to faces », présenté à European Conference on Visual Perception, Toulouse, France, 2011, p. 157.

A. Brilhault, M. A. Mathey, N. Jolmes, S. M. Crouzet, et S. J. Thorpe, « Saccades to Color: An Ultra-Fast Controllable Mechanism to Low-Level Features », présenté à Vision Science Society, Naples, Florida, 2011, vol. 11, p. 553-553.

N. Jolmes, A. Brilhault, M. Mathey, et S. J. Thorpe, « Ultra-rapid saccades to faces in complex natural scenes: A masking study », présenté à European Conference on Visual Perception, Toulouse, France, 2011, p. 96.

M. Mathey, A. Brilhault, N. Jolmes, et S. J. Thorpe, « Ultra-rapid saccades: Faces are the best stimuli », présenté à European Conference on Visual Perception, Toulouse, France, 2011, p. 155.

S. Kammoun, G. Parseihian, O. Gutierrez, A. Brilhault, A. Serpa, M. Raynal, B. Oriola, M. J.-M. Macé, M. Auvray, M. Denis, S. J. Thorpe, P. Truillet, B. F. G. Katz, et C. Jouffrais, « Navigation and space perception assistance for the visually impaired: The NAVIG project », *IRBM*, vol. 33, n° 2, p. 182-189, avr. 2012.

B. F. G. Katz, F. Dramas, G. Parseihian, O. Gutierrez, S. Kammoun, A. Brilhault, L. Brunet, M. Gallay, B. Oriola, M. Auvray, P. Truillet, M. Denis, S. Thorpe, et C. Jouffrais, « NAVIG: Guidance system for the visually impaired using virtual augmented reality », *Technology and Disability*, 2012.

B. F. G. Katz, S. Kammoun, G. Parseihian, O. Gutierrez, A. Brilhault, M. Auvray, P. Truillet, M. Denis, S. Thorpe, et C. Jouffrais, « NAVIG: Augmented reality guidance system for the visually impaired », *Virtual Reality*, 2012.

J. Borovec, O. Gutierrez, A. Brilhault, S. Kammoun, P. Truillet, et C. Jouffrais, « Fusion of heterogeneous data for better positioning of visually impaired pedestrians », *En cours de soumission*, 2014.